The Open University

M303

Further pure mathematics

History Reader

This publication forms part of an Open University module. Details of this and other Open University modules can be obtained from the Student Registration and Enquiry Service, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)845 300 6090; email general-enquiries@open.ac.uk).

Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.

To purchase a selection of Open University materials visit www.ouw.co.uk, or contact Open University Worldwide, Walton Hall, Milton Keynes MK7 6AA, United Kingdom for a brochure (tel. +44 (0)1908 858779; fax +44 (0)1908 858787; email ouw-customer-services@open.ac.uk).

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 2014.

Copyright © 2014 The Open University

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, transmitted or utilised in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without written permission from the publisher or a licence from the Copyright Licensing Agency Ltd. Details of such licences (for reprographic reproduction) may be obtained from the Copyright Licensing Agency Ltd, Saffron House, 6–10 Kirby Street, London EC1N 8TS (website www.cla.co.uk).

Open University materials may also be made available in electronic formats for use by students of the University. All rights, including copyright and related rights and database rights, in electronic materials and their contents are owned by or licensed to The Open University, or otherwise used by The Open University as permitted by applicable law.

In using electronic materials and their contents you agree that your use will be solely for the purposes of following an Open University course of study or otherwise as licensed by The Open University or its assigns.

Except as permitted above you undertake not to copy, store in any medium (including electronic storage or use in a website), distribute, transmit or retransmit, broadcast, modify or show in public such electronic materials in whole or in part without the prior written consent of The Open University or in accordance with the Copyright, Designs and Patents Act 1988.

Edited, designed and typeset by The Open University, using the Open University TEX System.

WEB 03766 2

1.1

Contents

1 The origins of structural algebra **4**
1.1 From the solution of equations to Galois Theory 4
1.2 From Galois Theory to finite group theory 19
1.3 Finite fields (Galois, Jordan, Steinitz) 23

2 The origins of topology and metric spaces **26**
2.1 The importance of Fourier series 26
2.2 Potential theory and the Dirichlet principle 38
2.3 Metric spaces – Fréchet 44
2.4 Topological spaces – Hausdorff 45

3 Biographies **48**
3.1 Pythagoras 48
3.2 Euclid 49
3.3 Marin Mersenne 49
3.4 Pierre de Fermat 50
3.5 Leonhard Euler 51
3.6 Joseph-Louis Lagrange 53
3.7 Adrien-Marie Legendre 53
3.8 Joseph Fourier 55
3.9 Carl Friedrich Gauss 56
3.10 Peter Gustav Lejeune Dirichlet 59
3.11 Bernard Riemann 60
3.12 Georg Cantor 60
3.13 Felix Hausdorff 61

References **62**

Acknowledgements **65**

1 The origins of structural algebra

1.1 From the solution of equations to Galois Theory

It can be hard to believe that modern structural algebra (groups, rings, fields, homomorphisms and isomorphisms) grew out of straightforward mathematical problems, and that the algebra we study at school is its legitimate ancestor. But such is the case, and it can be instructive to see why. There are, in fact, few more dramatic stories in all of mathematics than the way that attempts to solve polynomial equations turned into the study of groups and fields, as we will now describe.

Cubic and quartic equations

Mathematicians in ancient Mesopotamia some 4000 years ago knew how to solve problems we would write today as quadratic equations. Here is one of their surviving problems. (The form of the numbers has been modernised.)

From Robson (2007, p. 104).

> I took away my square-side from inside the area and it was 870. You put down 1, the projection. You break off half of 1. You combine 0.5 and 0.5. You add 0.25 to 870. 870.25 squares 29.5. You add 0.5 which you combined to 29.5 so that the square-side is 30.

The first line tells us that the problem is to find the side of a square when the area minus the side is 870. The modern version of this would be the equation $x^2 - x = 870$, which we are to solve for x. We can follow what is going on by looking at what we do to solve $ax^2 - bx - c = 0$. The method is to take 1 (b, the coefficient of x, so to speak), halve it and square it to give 0.25 (or $(b/2)^2$), add it to the 870 (which is ac in our equation), take the square root ($\sqrt{(b/2)^2 - ac}$), add 0.5 (which is $b/2$) and thus obtain the answer (the one we learn today as $\dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$). Notice that in the Mesopotamian problem c is negative. The Mesopotamian scribe naturally considered only the positive square root.

A perfectly general theory of quadratic equations having at least one positive real root was written down in the ninth century CE by the Islamic mathematician Al-Khwarizmi, but there the matter rested until, by the early sixteenth century, it had become something of a scandal that no progress had been made on solving cubic equations in terms of an algebraic formula of some kind. There was then a flurry of activity, spurred on by a very public rivalry between the leading contestants (as success at problem solving could lead to a lucrative contract as a scientific advisor at court), in which honours were finally divided between Niccolò Tartaglia and Gerolamo Cardano (Figure 1.1).

See the Foreword in Cardano (1543).

In those days mathematicians distinguished sharply between equations such as $x^3 + ax = b$ and $x^3 + b = ax$, because they required that the coefficients a and b be positive in each case. A mathematician in Bologna called Scipione del Ferro found a way of solving some cubic equations

(those of the form $x^3 + ax = b$, which have one positive real root – the others are complex and eluded del Ferro) and he confided it to a pupil, Mario Fior. Fior in turn challenged Tartaglia to a problem-solving competition, thinking that with his secret method he was bound to win, but the night before the competition Tartaglia found the way to solve Fior's cubics, and won. The fame he acquired brought him to the attention of Cardano, and Tartaglia confided his methods to Cardano under conditions of strict secrecy in exchange for an introduction to the local prince.

Figure 1.1 Gerolamo Cardano (1501–1576)

Cardano then found out how to solve all types of cubic equation, and his assistant, Ludovico Ferrari, showed how to solve equations of degree 4 (quartic equations). They chafed for some time under Cardano's promise because they had so much new to say, until they heard about Scipione del Ferro and Fior; they then decided that Tartaglia had not, after all, possessed a secret method, so it was permissible to publish what they had discovered. Tartaglia was predictably furious and a long row began between him and Ferrari. But the book that Cardano published, the *Ars Magna* (*Great Art*) of 1543, moved algebra out of the limited domain of equation-solving it had inhabited for centuries.

The general cubic equation

$$ax^3 + bx^2 + cx + d = 0$$

can always be written in a simpler form. We divide through by a to get an equation of the form

$$x^3 + b'x^2 + c'x + d' = 0.$$

If we replace x with $x - \frac{b'}{3}$ we get

$$(x - \tfrac{b'}{3})^3 + b'(x - \tfrac{b'}{3})^2 + c'(x - \tfrac{b'}{3}) + d' = 0.$$

The coefficient of x^2 is $-3\frac{b'}{3} + b' = 0$, so the equation now has this form:

$$x^3 + px + q = 0.$$

It is the equation in this form that we solve; it is then merely a tedious exercise to work back to the solution of the original equation.

Tartaglia, and Cardano after him, exploited the similarity in form between the above equation and

$$(u + v)^3 - 3uv(u + v) - (u^3 + v^3) = 0.$$

This is an algebraic identity, true for all values of u and v. So if we set $p = -3uv$ and $q = -(u^3 + v^3)$, it follows that x is $u + v$. It remains to solve this pair of simultaneous equations for the unknowns u and v, and the solutions turn out to be

$$u = \sqrt[3]{-\tfrac{q}{2} + \sqrt{(\tfrac{q}{2})^2 + (\tfrac{p}{3})^3}},$$

$$v = \sqrt[3]{-\tfrac{q}{2} - \sqrt{(\tfrac{q}{2})^2 + (\tfrac{p}{3})^3}}.$$

Notice that in each case we have three cube roots to consider: $\omega^k u$ and $\omega^m v,\ k, m = 0, 1, 2$, where ω is a primitive cube root of unity ($\omega \neq 1$, $\omega^3 = 1$). A little algebra shows that the three solutions of the cubic are

$$u + v,\ \omega u + \omega^2 v,\ \omega^2 u + \omega v.$$

Important though this achievement is, the formula is often of little value. Consider, as Cardano did, this cubic:

$$x^3 - 7x + 6 = 0.$$

It has three real roots $x = 1, 2, -3$, but in this case $(\frac{q}{2})^2 + (\frac{p}{3})^3 = -\frac{100}{27}$ is negative, so the formula involves complex numbers. But for cubic equations with only one real root and two complex roots, the formula gives the real root without involving complex numbers, so the occurrence of complex numbers is counter-intuitive. This is why cubic equations mark the point at which imaginary numbers forced themselves into mathematics. They did not come in with quadratic equations but with cubics, where the value of the roots is completely obscured unless an algebra is developed to handle these imaginary numbers.

Cardano's assistant Ludovico Ferrari seems to have been so violent at times that Cardano wrote in his biography of Ferrari that he was frightened of him. However, he turned out to be the mathematician who discovered how to solve the quartic equation, and he always acknowledged a lifelong debt to Cardano. The trick that we used to remove the quadratic term from a cubic equation can also be used to remove the cubic term from an equation of degree 4, so the equation can always be put in the form

See Cardano (1663, pp. 568–9) and Morley (1854, p. 266).

$$x^4 + px^2 + qx + r = 0.$$

Ferrari had the idea of putting x^4 on one side and all the other terms on the other side, then completing the square for the terms on the right-hand side. So he wrote

$$x^4 = -px^2 - qx - r,$$

and added $2zx^2 + z^2$ to each side. This gave him

$$x^4 + 2zx^2 + z^2 = (2z - p)x^2 - qx + (z^2 - r),$$

and he now looked for the value of z that makes the right-hand side a perfect square. This is found, by the $b^2 - 4ac$ test, to be a root of

$$q^2 = 4(2z - p)(z^2 - r).$$

This is a cubic in z and so, by Cardano's earlier work, it can be solved. Taking any value of z that is a solution of this equation, we can solve the corresponding equation for x^2 and take square roots to get the required values of x.

Quintic equations – Lagrange's work

The solution of quadratic, cubic and quartic equations involves the usual operations of arithmetic (addition, subtraction, multiplication and division) and the taking of square or cube roots. For this reason the solution method is called 'solving by radicals', from the Latin word *radix* for a root.

But progress halted with the quartic equation, and attempts by several mathematicians to solve the general quintic equation by radicals all failed. On the one hand, it was increasingly clear to mathematicians that every polynomial equation has as many roots as its degree indicates (this is the content of the fundamental theorem of algebra, which was only proved in the early nineteenth century but was long believed). On the other hand, no one could find an explicit formula involving nothing but addition, subtraction, multiplication, division and root extraction that exhibited the roots of a polynomial equation of degree 5 in terms of the coefficients of the equation.

The first person to take a fresh look at the problem was Joseph-Louis Lagrange. His plan was to investigate why cubic and quartic equations can be solved, and then either to see if the quintic can be solved for the same sort of reason, or to conclude that perhaps the quintic cannot in general be solved by radicals.

See Lagrange (1770).

The spur to Lagrange's thinking was this. A cubic equation $x^3 + px + q = 0$ has three roots; let us call them u, v, w. We have

$$u + v + w = 0,\ uv + vw + wu = p,\ uvw = -q.$$

Now, the expressions $u + v + w$, $uv + vw + wu$ and uvw are symmetric; they do not change their values under permutations of u, v and w. In fact, any symmetrical expression in u, v and w can be written in terms of these three. But the formula for the roots involves square roots and cube roots, and square roots take only two values and cube roots only three. The converse is also true. For example, an expression in u, v and w that takes only two values should be the root of a quadratic equation whose coefficients are rational functions of p and q.

Suppose σ_1 is such an expression, and under the six permutations of u, v and w its other value is σ_2. Consider the equation

$$x^2 - (\sigma_1 + \sigma_2)x + \sigma_1\sigma_2 = 0.$$

A permutation either leaves σ_1 and σ_2 alone or exchanges their values, so any permutation of u, v and w leaves $\sigma_1 + \sigma_2$ and $\sigma_1\sigma_2$ alone. But that means that $\sigma_1 + \sigma_2$ and $\sigma_1\sigma_2$ are expressible in terms of p and q.

A very similar argument deals with expressions in u, v and w that take only three values. So Lagrange speculated that the cubic equation is solvable because there are expressions in u, v and w that take only two or three values.

In this case, the key expression is $\sigma = (u - v)(v - w)(w - u)$, which takes just two values under the six permutations: σ and $-\sigma$. This means that σ^2 is symmetric, so it is an expression in p and q. Since p is of degree 2 in u and v, q is of degree 3 and σ^2 is of degree 6, this suggests trying expressions in p^3 and q^2, which recalls the formula for the roots. A little algebra reveals that

$$\sigma^2 = -4p^3 - 27q^2,$$

from which we deduce that

$$\pm\frac{\sigma}{6\sqrt{-3}} = \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}.$$

So, to explain why the cubic equation is solvable by radicals, Lagrange argued that it is the existence of an expression in the roots that takes only two values that accounts for the existence of the square root in the formula, and then the solution is completed by nothing more elaborate than taking a pure cube root.

We can now sketch Lagrange's very lengthy analysis of the quartic equation

$$X^4 + aX^3 + bX^2 + cX + d = 0.$$

Let the roots of this equation be x_1, x_2, x_3 and x_4. From Ferrari's work, it is clear that the solution involves solving a cubic equation and then taking square roots of expressions that involve the roots of that cubic equation. From his analysis of the general cubic equation, Lagrange knew that the fact that the quartic is solvable by radicals is probably due to the existence of expressions in the roots that take fewer than 24 values as the roots are permuted, and indeed take only 2 or 3 values, which would mean they were roots of quadratic or cubic equations.

Compare Book B, Chapter 5, Exercise 5.4.

Lagrange looked at all the known methods for solving the quartic equation, and showed that they all depended on the existence of partially symmetric functions of the roots – for example, the expression $x_1x_2 + x_3x_4$, which takes only three different values, or the expression $x_1 + x_2 - x_3 - x_4$, which takes only six. Towards the end of a long memoir in 1770 he concluded the following (Lagrange, 1770, p. 304):

> We have not only brought together these different methods and shown their connection and interdependence, but we have also – and this is the main point – given the reason, a priori, why some of them lead to resolvents of degree three and others to resolvents of degree six that can be reduced to degree three. This is due, in general, to the fact that the roots of these resolvents are functions of the quantities x_1, x_2, x_3, x_4, which, like the function $x_1x_2 + x_3x_4$, take on just three values under all the permutations of these four quantities; or which, like the function $x_1 + x_2 - x_3 - x_4$, take on six values that are pairwise equal but have opposite signs; or which take on six values that can be separated into three pairs such that the sum or product of the values in each pair is not changed by the permutations of the quantities x_1, x_2, x_3, x_4 .... The general solution of quartics depends uniquely on the existence of such functions.

Accordingly, when it came to the quintic, Lagrange looked for expressions in five symbols x_1, x_2, x_3, x_4, x_5 that took only three or four values as the xs were permuted. Such expressions would satisfy equations of degree 3 or 4, which he called 'resolvents', and these equations would have coefficients that were explicit combinations of the coefficients of the quintic equation, and so could be solved. However, because Lagrange worked with expressions directly, and did not make a separate study of permutations and their combinations, his analysis was very clumsy and long-winded.

This may have contributed to his eventual opinion of the difficulty of the problem of solving the quintic, which was not that it could not be done but that success was unlikely. He also noted that even the solvability of the quartic depended on a happy accident: the expression $x_1 + x_2 - x_3 - x_4$ that takes six values in fact takes three values and their negatives, so it leads not to a typical equation of degree 6 but one that can be solved by solving a quadratic and a cubic equation.

The best that Lagrange could do was to find an expression in the roots that took only six different values when the roots were permuted. But this led to an equation of degree 6, and there seemed to be no reason why it would be one of those equations that could be solved by solving first a quadratic and then a cubic. If it could not, there was no hope of solving the general equation of degree 5 by tackling an equation of higher degree. He expressed his downbeat conclusions this way (Lagrange, 1770, p. 403):

> Here, if I am not mistaken, are the true principles for the solution of equations and for the best analysis of its conduct. As one can see, everything reduces to a kind of combinatorial calculus by which one finds a priori the results to be expected. It may seem reasonable to apply these principles to equations of the fifth and higher degrees, of which the solution is at present unknown, but such application requires too large a number of examinations and combinations whose success is very much in doubt. Thus, for the time being, we must give up this work. We hope to be able to return to it at another time. At this point we shall be satisfied with having set forth the foundations of a theory which appears to us to be new and general.

In fact, Lagrange never returned to the topic and his next long paper on the topic of solving equations was on the method of finding numerical approximations to their roots, which some historians have suggested is further evidence that he despaired of solving the quintic algebraically. In fact, such an approach is consistent with the belief that even if the equations could be solved by radicals, the solution expressions would be unintelligible, as the example of the cubic equation has already suggested.

Little was then done for a further 30 years, until the Italian mathematician Paolo Ruffini announced that the general quintic equation was not solvable by radicals. He was unlucky in not being sympathetically read, for his argument is flawed at only one point and his conclusion is correct. The first fully correct proof was then given in 1824 by the young Norwegian mathematician Niels Henrik Abel (Figure 1.2). Each tackled the problem of what the succession of a sequence of equations of low degree would look like, which is to say they wrote down expressions that are pure n_1th roots of the coefficients of the quintic equation, then expressions that are pure n_2th roots of the coefficients of the quintic equation and the n_1th roots already obtained, and so on. They considered what properties such expressions must have if they contribute to the solution of the quintic, and each showed, Abel more rigorously than Ruffini, that on symmetry grounds no expressions exist that correspond to the roots of the general quintic.

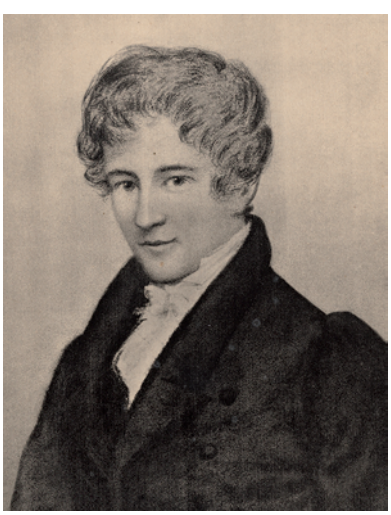

Figure 1.2 Niels Henrik Abel (1802–1829)

A good reference for all this material is Tignol (2001).

Quintic equations – Galois's work

This material relates to Book B, especially Chapters 5 and 7.

Figure 1.3 Augustin-Louis Cauchy (1789–1857)

A crucial breakthrough that Abel exploited was Cauchy's study of permutations. Augustin-Louis Cauchy (Figure 1.3) was the dominant French mathematician of his day, a gifted if egocentric thinker with the ability to work on any and every branch of mathematics and switch quickly between them, who is probably best remembered for his work in creating rigorous real and complex analysis.

In a long paper of 1815, Cauchy had shown that the number of values a rational expression in n variables can take must be greater than or equal to the largest odd prime that divides n unless it is equal to 2. In only slightly more modern language, Cauchy showed that the size of an orbit of a subgroup of the group S_n is either greater than or equal to p, the largest odd prime dividing n, or it is 2. The size of the orbit is the index of the stabiliser of a point of that orbit, and so the size of the stabiliser is either less than or equal to $\frac{n!}{p}$ or it is $\frac{n!}{2}$. This rules out subgroups of index 3 or 4 in S_5, so it has no subgroups of orders 40 or 30.

This connects neatly to what has been done before, because an expression in n variables that takes precisely k values under the $n!$ permutations of the variables is one that is fixed by a subgroup of order $\frac{n!}{k}$ of the full permutation group. So the non-existence of a group of order 30 in the permutation group on five symbols corresponds to the non-existence of an expression in five variables that takes on $\frac{120}{30} = 4$ values. Similarly, there is no expression taking only $\frac{120}{40} = 3$ values. This in turn means that the roots one can extract when solving a quintic equation are either square roots or fifth roots, and this was the decisive simplification that enabled Abel to do what Lagrange had not been able to do.

The problem now became identifying which polynomial equations of degree 5 or more are solvable by radicals. There are many of these. Carl Friedrich Gauss – see Subsection 3.9 – had given a rich theory of the equations that give n roots of unity:

$$x^n - 1 = (x-1)(x^{n-1} + x^{n-2} + \cdots + x + 1) = 0,$$

of which the interesting one is

$$x^{n-1} + x^{n-2} + \cdots + x + 1 = 0.$$

He had shown, for example, that when $n = 19$ the equation is solved by solving a succession of one quadratic and two cubic equations (note that $2 \cdot 3 \cdot 3 = 18$). When $n = 17$, and so $n - 1 = 16 = 2^4$, the equation is solved by solving a sequence of quadratics, and so by straightedge and circle (unmarked ruler and compass) alone – the first new construction of that kind since the Greeks. Gauss was only 18 when he discovered this result in 1796.

The particular case of constructions that are possible by ruler and compass alone is discussed in Book E, Chapter 19, Section 2.

The problem of deciding which polynomials are solvable by radicals is the one solved by Évariste Galois (Figure 1.4), a French mathematician who died at the age of 20 as a result of injuries obtained in a duel.

Galois made several innovations in his work. First, he observed that although every polynomial equation has a group that permutes its roots,

The definitive edition of Galois's works is Neumann (2011), which gives the French texts with a facing English translation. For a short, accurate life of Galois, see the essay about him in Rothman (1989).

different equations of a given degree (say n) may have different groups. Consider, for example, the equation for complex fifth roots of unity. It is

$$x^4 + x^2 + x^2 + x + 1 = 0. \tag{1.1}$$

Let $\omega = e^{2\pi i/5}$ be one of the roots; then the other roots are ω^2, ω^3 and ω^4. In this instance, every permutation of the roots is a power of the permutation ρ that takes ω to ω^2:

$$\rho(\omega) = \omega^2.$$

To see this, observe first that every permutation of these roots is known when it is known on ω. Next, we know that

$$\rho(\omega^2) = (\omega^2)^2 = \omega^4, \quad \rho(\omega^3) = (\omega^3)^2 = \omega^6 = \omega$$
$$\text{and} \quad \rho(\omega^4) = (\omega^4)^2 = \omega^8 = \omega^3,$$

because $\omega^5 = 1$. So we can write ρ in cycle notation as $(\omega^1, \omega^2, \omega^4, \omega^3)$. Now observe that ρ^2 can be written in cycle notation as $(\omega^1, \omega^4)(\omega^2, \omega^3)$, and that ρ^3 can be written in cycle notation as $(\omega^1, \omega^3, \omega^4, \omega^2)$, which is also ρ^{-1}. So this equation has as the group of permutations of its roots the cyclic group of order 4, not the full permutation group on four objects, which has order 24. This example was known to Gauss and Abel.

Figure 1.4 Évariste Galois (1811–1832)

Galois then adapted Lagrange's ideas to the new, more precise setting, and to state what he did we need a definition. We say that a function is *rationally known* if its numerical value is a rational function of the coefficients of the given equation and any quantities so far adjoined. Note that the decision as to whether a given expression is rationally known may change as quantities are adjoined to the field; it depends on where we are in the process of solving the equation.

When a polynomial equation is solved and its roots adjoined to the field in which the coefficients live, we obtain a new, larger field that contains the original field, and we can consider the permutations of the roots that leave the elements of the original field fixed. These permutations form a group, and we have just seen that for equation (1.1) this group is a cyclic group of order 4. But that equation is obviously special. For a typical or 'general' equation of order n with n distinct roots, one expects that the group that permutes the roots will be the full permutation group on n elements. In this case, we can also argue that any function that is rationally known is a rational function of the coefficients of the equation and, as such, is a rational function of the symmetric functions and therefore invariant under all permutations.

One key insight that Galois had was that the way to consider what happens in the process of solving a polynomial equation is to think about the group of permutations of its roots, and to watch what happens as quantities are adjoined to the original field in which the coefficients live. Different equations may have different groups, and a complete understanding of these groups would give a complete understanding of the equations themselves. This was a lesson that it took the mathematical community more than a generation to learn.

Galois now proved the following theorem.

Theorem 1.1 *Galois's First Theorem*

Given an equation (to be called the given or original equation) with roots $x_1, \ldots, x_n$, there is a group G permuting the roots of this equation with the property that

- every G-invariant function is rationally known
- every rationally known function is G-invariant.

Next Galois considered what happens when an equation is solved by solving a succession of equations of lower degree. If you look back at the solution of the cubic equation, you will see that we start with $x^3 + px + q = 0$ and proceed in stages as follows. It is best to assume that we have already adjoined a cube root of unity to the field of rational numbers (this is a number ω such that $\omega^3 = 1$).

1. We calculate $\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3$, which is a rational number.
2. We calculate $\sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}$, which we do not expect to be a rational number. We'll call one of the resulting square roots σ and the other one $-\sigma$. We are going to use σ, so we now say it is rationally known, meaning that we are now working in the field of the rational numbers with σ adjoined.
3. We calculate $-\frac{q}{2} \pm \sigma$, which does not take us to a larger field.
4. We calculate $u = \sqrt[3]{-\frac{q}{2} + \sigma}$ and $v = \sqrt[3]{-\frac{q}{2} - \sigma}$, as well as $\omega u, \omega^2 u, \omega v$ and $\omega^2 v$, where ω is a cube root of unity. This takes us from the field of the rational numbers with σ adjoined to a bigger field with these cube roots adjoined.
5. We observe that elements of this bigger field, namely $u + v$, $\omega u + \omega^2 v$ and $\omega^2 u + \omega v$, solve the cubic equation.

So, said Galois, solving an equation by radicals is a matter of steadily enlarging the field of the coefficients by adjoining pure roots (square roots, cube roots and, generally, nth roots). As the field of quantities that are rationally known is enlarged by each successive adjunction, one looks to see what the group of the equation now is. One can also be more imaginative and imagine adjoining roots of other equations, which he called auxiliary equations, and looking to see what happens.

To state what happens, Galois claimed the following theorem, which unfortunately illustrates very clearly how badly he explained his ideas (Neumann, 2011, p. 119).

Theorem 1.2

If one adjoins to a given equation the root r of an irreducible auxiliary equation, one of two things will happen: either the group of the equation will not be changed, or it will be partitioned into p groups each belonging respectively to the proposed equation when one adjoins to it each of the roots of the auxiliary equation; these groups will enjoy the remarkable property that one will pass from one to another by operating on all the permutations of the first with one and the same substitution of letters.

His idea is that if you regard a new number r as known then either you are no better off, or you are. You are better off if the group of symmetries of the equation shrinks, in the same way that Lagrange thought he was getting nearer to finding a solution every time he found an expression with an unexpectedly small number of symmetries.

Galois called an arrangement of n symbols a *permutation*, and a rearrangement of them (what we would call a permutation) a *substitution*. Such, eventually, was the success of Galois's ideas that finite groups were for a long time called substitution groups. But his terminology is the least of our problems. Doubtless because of the circumstances in which Galois wrote the memoir he was working on, he struck out the statement that the number p is the degree of the auxiliary equation, and failed to state that the partition is into j groups, where j divides p. Moreover, here 'group' simply means subset: one gets a subgroup of the original group (here called H) and its cosets. All the elements (Galois wrote *permutations*) in each coset are obtained from the first one by a single multiplication – this is how we obtain gH from H.

He did, however, sketch a proof, which is more than can be said for the rest of his memoir. He went on to state the next theorem as follows.

Theorem 1.3

If all the roots of an auxiliary equation are adjoined, each of the above 'groups' has the same elements.

Here he could only note 'One will find the proof'. The next theorem was as follows.

Theorem 1.4

Adjoining the *numerical* value of a certain function of the roots reduces the group of the equation to those elements that leave this function invariant.

Galois now came to his fundamental question: when is an equation solvable by simple radicals? By this he meant, by adjoining a sequence of nth roots for various n.

Certainly, an equation solvable by simple radicals is such that the group can be diminished to a group with a single element by adjoining a suitable sequence of roots of prime degree. Conversely, Galois said, if the group breaks up in this fashion then the equation is solvable by simple radicals. The reason is that a simple radical is a root of an equation of the form $x^p - a = 0$, where a is rationally known, and the group of permutations of the roots of the equation $x^p - a = 0$ has a particular structure that in turn forces a structure on the group of all permutations of the roots of the original equation.

Galois did try to explain this by indicating how this analysis held up in the case of quartic equations. The 'group' of the equation contains 24 *substitutions* (in our language, the 24 permutations of four objects) and, on adjoining a (suitable) square root, decomposes into two 'groups', each of 12 *substitutions*. Galois listed the elements of one of these 'groups': it is identical with what we call A_4. This group splits into three 'groups', in line with Galois's theorems on adjoining a radical of the third degree (a cube root). One of these is the group that came to be called the Klein group, $V = \{e, (12)(34), (13)(24), (14)(23)\}$, and it splits into two 'groups': $\{e, (12)(34)\}$ and $\{(13)(24), (14)(23)\}$ – notice that this second 'group' is really just a coset of the first. Adjoining a square root isolates the 'group' $\{e, (12)(34)\}$, which is reduced to the identity by adjoining a square root.

Galois concluded his memoir by investigating irreducible equations of prime degree, p. Recall that the process of solving an equation by radicals reduces the group of the equation in successive steps. Galois's analysis now proceeded by working back up from the trivial group (corresponding to the complete factorisation of the equation) to the original group.

Consider the group of an irreducible equation of prime degree p that is solvable by radicals. Galois showed that either the equation is unaffected by the adjunction of a radical or it is reduced completely. Therefore, in the reduction process, the last group reached before the group with only one element must be a group of p elements, and (as Cauchy had shown) this means the group must be the cyclic group of order p.

What about the next group up (G, say)? Galois showed that it must be made up of a certain number of 'groups' having the same 'substitutions' as the cyclic group of order p (which makes sense if we allow that Galois is alluding to the corresponding coset decomposition – the same transformations (substitutions) are applied to different lists (or permutations) of the roots). From this, by an ingenious argument, Galois deduced that the next group in the decomposition must consist of elements of the form

$$x_k \mapsto x_{ak+b},\ a, b \in \{0, 1, 2, \ldots, p-1\} \text{ modulo } p.$$

Then Galois showed that the whole group of the equation must be the group G just found, and therefore that a polynomial equation of prime degree is solvable by radicals if, and only if, its group is of the form of G.

The crucial property of this group is that it acts transitively on the roots. Moreover, no substitution of the form $x_k \mapsto x_{ak+b}$ other than the identity fixes two elements, so once two roots are known rationally so are the rest, and the group of the equation reduces to the group with one element. Therefore, for the given equation to be solvable by radicals, it is necessary and sufficient that once any two roots are given, all the others are rationally known.

In the case of a quintic equation, Galois's result implies that the largest 'group' of an equation solvable by radicals is the group of order 20 consisting of the elements

$$x_k \mapsto x_{ak+b},$$

where a and b are integers modulo 5 and b is not congruent to zero. You might like to show that in this case G is generated by $(1, 2, 3, 4, 5)$ and $(2, 3, 5, 4)$.

Therefore the general quintic equation, which has group S_5, is not solvable by radicals.

Galois's letter to Chevalier

The night before Galois went out to fight his duel, fearing that he would be killed, he wrote a hasty note to his friend Auguste Chevalier. This, as you will see, is the first place where he drew attention to the importance of adjoining roots that reduce the group of the equation to a normal subgroup. He wrote (Neumann, 2011, p. 85):

> According to Propositions II and III of the first memoir one sees a great difference between adjoining to an equation one of the roots of an auxiliary equation or adjoining them all.
>
> In both cases the group of the equation is partitioned by the adjunction into groups such that one passes from one to another by one and the same substitution; but the condition that these groups should have the same substitutions does not necessarily hold except in the second case. That is called a *proper decomposition*.
>
> In other words, when a group G contains another H, the group G can be partitioned into groups each of which is obtained by operating on the permutations of H with one and the same substitution, so that
>
> $$G = H + HS + HS' + \cdots .$$
>
> And also it can be decomposed into groups all of which have the same substitutions, so that
>
> $$G = H + TH + T'H + \cdots .$$

Here Galois presented the left and right coset decompositions of a group G with respect to a subgroup H. He did not say that two cosets are either the same or distinct, but he immediately went on to make the crucial

observation that the important case is when the left and right coset decompositions agree: every right coset is a left coset and every left coset a right coset. His word for this was 'proper', and it did not last. We say that when we have such a decomposition, the subgroup H is a normal subgroup of G. In his words:

> These two kinds of decomposition do not ordinarily coincide. When they coincide the decomposition is said to be *proper*.
>
> It is easy to see that when the group of an equation is not susceptible of any proper decomposition one may transform the equation at will, and the groups of the transformed equations will always have the same number of permutations.

This is not at all easy to see, but the key point that the behaviour of the group of the equation will answer all your questions about how to solve the equation itself is clear. Galois went on (Neumann, 2011, p. 87):

> When, on the contrary, the group of an equation is susceptible of a proper decomposition, so that it is partitioned into M groups of N permutations, one will be able to solve the given equation by means of two equations: the one will have a group of M permutations, the other one of N permutations.

Here Galois's confusion of subgroup and coset gets in the way. If we look back at the treatment of the general cubic equation, we can see that we started with the field of rational numbers, $\mathbb{Q}$, to which we had adjoined a cube root of unity, ω – which gives us the field $\mathbb{Q}(\omega)$. Then we first solved a quadratic equation by taking a pure square root and adjoining σ, and then solved two pure cubic equations by taking their cube roots and adjoining what we called u and v. The resulting field contains the solutions of the cubic equation. So we have formed a succession of fields

$$\mathbb{Q}(\omega) < \mathbb{Q}(\omega)(\sigma) < \mathbb{Q}(\omega)(\sigma)(u, v).$$

What Galois said in this case was that there is a group of six elements that permutes the roots of this cubic, but it breaks up into a subgroup, H, of three elements and the coset of this subgroup in the bigger group. In this setting, Galois's numbers M and N become $M = 2$ and $N = 3$, and his two 'groups' are a group of order 3 and its coset. The group H permutes the elements of the field $\mathbb{Q}(\omega)(\sigma)(u, v)$ and fixes those in the field $\mathbb{Q}(\omega)(\sigma)$. There is also a group of order 2 that permutes the elements of the field $\mathbb{Q}(\omega)(\sigma)$ and fixes the elements of the field $\mathbb{Q}(\omega)$ – in other words, the group that switches $+\sigma$ and $-\sigma$.

Galois continued:

> Therefore once one has effected on the group of an equation all possible proper decompositions on this group, one will arrive at groups which one will be able to transform, but in which the number of permutations will always be the same.
>
> If each of these groups has a prime number of permutations the equation will be soluble by radicals; if not, not.

In the case of the cubic, these numbers are 2 and 3. In the case of the quartic, these numbers are 2, 3, 2 and 2. In each case, the equation is solvable by radicals.

Galois then observed:

> The smallest number of permutations which can have an indecomposable group, when this number is not prime, is $5 \cdot 4 \cdot 3$.

In other words, the permutation group A_5 is simple; it has no non-trivial normal subgroups. It's a remarkable thing to have noticed, and Galois gave no hint of his proof.

Cauchy's theory of permutation groups

Cauchy's work was done in two phases, the first between 1812 and 1815 and the second around 1844–45.

In his paper 'Mémoire sur le nombre, etc.' (1815), Cauchy investigated how many different values a rational expression in n variables can take when its variables are permuted in every possible way. He noted that Lagrange and other writers had shown that it was not always possible to find a polynomial that took a specified number of distinct values (not even a number of values that divided $n!$, which is obviously necessary). Ruffini had shown in 1799 that it is not possible to find a function of $n = 5$ variables that takes only four values. Cauchy now ruled out the possibilities that a function of 5 variables takes only three or only four values. We have already seen that his work in the first phase helped Ruffini and Abel to prove that the general quintic equation is not solvable by radicals. Now we look at it in a little more detail.

In a much later paper (1844) Cauchy returned to the theme, for reasons that one speculates are to do with a revival of public interest in Galois's work, and treated the subject much more systematically. He called a list of the n symbols in an order an 'arrangement' of n symbols, and a transition from one arrangement to another he called a 'substitution' or permutation. He introduced the permutation notation (x, y, z), which we call the cyclic notation for permutations, for the arrangement, considered the product of two permutations and noted that forming the product of two permutations is not necessarily commutative. He looked at powers of a permutation, defined the identity permutation (written 1 and which he called a unity), and defined the degree or order of a permutation (the lowest positive power of it that is equal to 1). He then introduced the idea of similar permutations (which we call 'conjugate', but Cauchy reserved that word for another concept). Next he introduced the idea of a group defined (we would say 'generated') by a set of permutations – which he called a system of *conjugate* substitutions – and proved that (Cauchy, 1844, p. 207):

> The order of a system of conjugate substitutions in n variables is always a divisor of the number N of arrangements that one can make with these variables.

Here $N = n!$. Cauchy also proved that the order of any element of this group divides the order of the group.

Cauchy was aware of the concept of a normal subgroup, although he did not refer to Galois's proper decompositions, and in his paper of 1844 he established a partial converse to Lagrange's Theorem, showing that if a prime p divides the order of a group then there is an element of order p in that group and so a subgroup of order p (the condition that p be prime is necessary: there is no subgroup of order 6 in A_4, although 6 divides 12, which is the order of A_4).

Wantzel's resolution of two impossible classical problems

We discuss these theorems in Book E, Chapter 19, Section 2.

We now look at the discovery of the (nowadays) celebrated impossibility theorems that follow from Galois Theory: it is impossible by straightedge and compass to trisect an arbitrary angle or to find the side of a cube that has double the volume of a given cube. These are two of the so-called three classical problems of Greek geometry. The third is the problem of squaring the circle: find by straightedge and compass alone a square that is equal in area to a given disc (i.e. construct a segment of length $\sqrt{\pi}$). It is deeper and different and cannot be discussed here – to this day, squaring the circle is a synonym for impossibility.

The three classical problems have a complicated history. On the one hand, they were not distinguished in classical times as *the* three outstanding problems. Another, for example, asked for the construction of a regular heptagon. Nor, on the other hand, is it clear what made them difficult – or, indeed, left them unsolved. It was known by the time of Euclid that the cube can be duplicated by a construction involving the intersection of a hyperbola and a parabola. A construction for trisecting an arbitrary angle was found by Archimedes that involves sliding a marked straightedge around while ensuring that it passes through a given point. We have no source that specifies a restriction to the use of a straightedge and compass alone; the implication seems to have been drawn in modern times from the absence of any other kind of 'instrument' in Euclid's *Elements*. Nor is it clear whether the problems ask for a proof that there is an angle one-third the size of a given one (or that there is a cube twice the size of a given one) or only for a precise constructive procedure.

It is not difficult to persuade yourself that a straightedge and compass construction can only involve rational operations on known lengths and the extraction of square roots. Duplicating the cube and trisecting the angle involve cubic equations and cube roots, and the consensus by 1800 was that straightedge and compass constructions will be inadequate to the task. That answer is correct, and the first mathematician to establish this result was a French mathematician, Pierre Laurent Wantzel, who published his demonstration in 1837, when he was 23, in a mainstream journal, the *Journal des mathématiques*. Wantzel promptly vanished into obscurity, however, from which he has only just been rescued by the Danish historian of mathematics Jesper Lützen (see Lützen, 2009). Prior to Lützen's work there were very few references. There was one earlier mention, by Julius Petersen (a Danish mathematician) in 1877, but no one picked it up except Felix Klein – an influential mathematician of the late nineteenth century after whom the Klein group is named – and he dropped the reference to

See Wantzel (1837).

Wantzel by name. The first historian of mathematics to mention Wantzel by name was the North American historian Florian Cajori in 1918.

Wantzel's paper was in four parts.

1. He formulated the geometric straightedge and compass procedure as an algebraic one.
2. He deduced that the solution of a geometric problem under this restriction appears as a root of an equation of degree 2^n.
3. He claimed that this equation is irreducible (his argument was inadequate, but Pedersen and Klein repaired it easily enough).
4. However, he went on, duplicating a cube and trisecting an angle lead in general to irreducible cubic equations, which therefore cannot be solved by a sequence of quadratics.

Why did Wantzel not become famous? Lützen argues that it may be that people thought the result was already known, because Gauss had claimed in the *Disquisitiones Arithmeticae* (1801) that the only polygons that are constructible by straightedge and compass have $2^n p_1 p_2 \cdots$ sides, where the p_i are distinct Fermat primes. Therefore the regular 9-gon is not constructible; therefore the angle of 40° that appears at the centre of the 9-gon is not constructible; this is one-third of the (constructible) angle of 120°; therefore angles cannot, in general, be trisected by ruler and compass. Admittedly, for once Gauss had not supported his claim with a proof, but no one seems to have doubted that Gauss knew how to prove it – so for anyone persuaded of Gauss's claim, Wantzel was saying nothing new except for providing a proof.

Lützen also argues that the result may not have seemed interesting. People of this opinion could argue that it merely confirmed what had been suspected all along: two of the classical problems are, after all, insoluble. The result leads nowhere, and on the other hand, the equations are indisputably solvable – why care about a requirement that they be solved in a particular and frankly old-fashioned way? It would seem that the taste for impossibility theorems came later.

Lützen could find only one article about Wantzel written by someone who knew him, from which we learn that Wantzel did not fulfil his early promise. Ground down by a heavy teaching load, which he attempted to soften with coffee and opium, he died at the age of 33, his work already forgotten.

1.2 From Galois Theory to finite group theory

Galois's theory was considered very difficult in its day, and his memoir on which equations are solvable by radicals was not even published by the time he died in 1832. Then, in 1843, the mathematician Joseph Liouville announced to the Paris Academy of Sciences that he had discovered among Galois's unpublished papers a 'solution which is both correct and profound' to the problem of determining when an irreducible equation of prime degree is solvable by radicals. He also promised to publish the

memoir together with a commentary elucidating it, and explained that Galois had entrusted his manuscripts to his friend Auguste Chevalier, who had prepared an annotated version of the memoir that he had recently brought to Liouville's attention. By this time Liouville was the editor of the *Journal des mathématiques*, and the hope was that he would see that Galois's ideas finally made it into print. Liouville did indeed prepare them for publication in his journal in December 1843, only to withdraw the paper at the last minute. His reasons remain unclear, but when in 1846 he published a longer account, including a fragment of another paper, the last letter to Chevalier and all of Galois's previously published papers, he notably failed to provide a commentary.

The lack of a commentary is even more suggestive because in the winter of 1843–44 Liouville had given a short course of private lectures in Paris on Galois's ideas, which the mathematician Joseph Serret attended. Unpublished notes by Liouville from this period show that he had filled in the largest gaps in Galois's sketchy exposition, but it seems he never pulled them together into a systematic exposition. Nor was Serret to do much better. He avoided the topic in the first edition of his own *Cours d'algèbre supèrieure* (Course of higher algebra) of 1849, perhaps politely leaving room for Liouville to deliver on his promise. In the second edition (1854), all he did with Galois's work was to summarise it in a few lines (on p. 4). He asked the question 'When is an equation solvable by radicals?', and replied:

> This difficult question was completely answered, at least for irreducible equations of prime degree, by Evariste Gallois [*sic*], former student of the École Normale and one of the deepest geometers [mathematicians] that France has produced. In a paper presented to the Academy of Sciences in 1831 and published in 1846 thanks to the efforts of Liouville, Gallois in effect proved the following beautiful theorem: For an irreducible equation of prime degree to be solvable by radicals it is necessary and sufficient that, given any two of the roots, the others are rationally deducible.

What Serret did contribute was a deepening appreciation of permutation group theory, which Cauchy had pioneered in 1815 and more substantially in 1845. And when, in the third edition (1866), Serret worked through Galois's memoir, he stuck closely to what Galois had done and how he had done it, noting that 'I have followed the order of the propositions adopted by Galois, but, very frequently, I had to close the gaps in the proofs' (*Cours*, vol. 2, p. 607). He did not present the reworking of Galois's propositions that was needed to give them a new life some 30 years after their author's death.

The man who did that was a French mathematician of the next generation, Camille Jordan (Figure 1.5). He wrote a 660-page book, the *Traité des substitutions et des équations algébriques* (1870), on the theory of permutation groups and polynomial theory of equations, which is a landmark in the creation of group theory.

Jordan's *Traité* is an attempt to establish a new field of algebra, and Galois Theory as Jordan presented it is only one part of that endeavour – he also applied group theory to a wide range of complicated geometrical situations in the *Traité*. Jordan also moved the approach away from solving equations and towards their associated groups. That said, his account provides the right template for putting Galois's fragmentary account together.

Figure 1.5 Camille Jordan (1838–1922)

Jordan followed Galois in associating to each polynomial equation a group, which he called the *Galois group* of the equation. He then showed, much as Galois had indicated, how the step-by-step process of solving an equation by successive adjunctions of roots of other polynomial equations can be understood as a series of operations on the Galois group. Moreover, Jordan had no method of finding the Galois group of an arbitrary equation except in the simplest cases, and this had the effect of making the journey from equations to groups something of a change of subject. It implied that mathematicians should loosen their grip on equations and take hold of groups instead.

Jordan's systematic theory of permutation groups was indeed highly abstract. Among the concepts he stressed, some of which he introduced, were the properties of commutativity, conjugacy, centralisers, transitivity, 'normal' subgroups (and, if only obliquely, quotient groups), group homomorphisms and isomorphisms. He was one of the first to use the Sylow Theorems in the treatment of finite groups.

The *Traité* is concerned only with finite groups, which Jordan often presented as matrix groups with entries modulo a prime number, and so it opens with some material on modular arithmetic. It then sets out the basic elementary propositions in the theory of groups: Lagrange's Theorem and Cauchy's Theorem. Lagrange's Theorem is stated for the first time in the form it retains to this day: the order of a subgroup divides the order of the group, and the proof is the coset decomposition argument. Cauchy's Theorem (**S**40 of his 1844 paper) is the partial converse: if a prime p divides the order of a group, there is an element of order p in the group.

Modular arithmetic was introduced in Book A, Chapter 3, Section 1.

Lagrange's Theorem for groups is proved in Book B, Chapter 5, Subsection 2.2.

Cauchy's Theorem is proved in Book B, Chapter 6, Subsection 3.3.

Jordan was very interested in transitive group actions: a group is n-fold transitive if, given two ordered sets of n distinct elements, there is an element of the group mapping the first set onto the other. Jordan proved the Orbit–Stabiliser Theorem: given a set of n elements of a group G, if these elements can be sent to m different places in G, and n is the order of the subgroup that leaves these elements fixed, then the order of the group is mn.

The Orbit–Stabiliser Theorem is proved in Book B, Chapter 6, Subsection 3.1.

Transitive group actions are important in the theory of polynomial equations because, as Jordan proved, the Galois group of a polynomial equation acts transitively on the roots if, and only if, the equation is irreducible. Accordingly, when one solves a polynomial equation by successive adjunction of numbers, the worthwhile adjunctions are the ones that reduce the size of the Galois group; every time a Galois group is reduced to one that acts intransitively on the roots, the polynomial equation factorises. Eventually the Galois group is reduced to the identity, the equation is completely factorised, and all the roots are therefore known.

So, in the study of irreducible quartic and quintic equations, only the transitive subgroups of S_4 and S_5 need to be considered. The Orbit–Stabiliser Theorem tells us that if a group acts transitively on a set of n elements then it has an orbit of order n, and therefore n divides the order of the group. In the case of subgroups of S_5 this tells us that a transitive subgroup contains a 5-cycle, whereas in the case of S_4 it tells us that the transitive subgroups have order $4m$ – but this allows for the possibility that we might be looking at the Klein group, which does act transitively but does not have an element of order 4.

The transitive subgroups of S_4 are $S_4, A_4, S(square)$ and V (the Klein group).

The transitive subgroups of S_5 are $S_5, A_5, G(20), S(pentagon)$ and C_5.

Next, Jordan looked in more detail at how a group breaks up into subgroups, a topic that Galois had hinted at very obscurely. Jordan was interested in a subgroup H of a group G that commuted with all the elements of G – a relationship he called 'permutable', meaning $gH = Hg$ for all g in G. (In terms of the elements involved, this says that for every $h \in H$ and every $g \in G$ there is an $h' \in H$ such that $gh = h'g$.) Jordan called G composite if it had a non-trivial subgroup of this kind and simple otherwise, and he observed that every group admits a chain of subgroups $G, G_1, G_2, \ldots$ in which G_1 is permutable with G, G_2 is permutable with G_1, and so on. He did not introduce a special word for a subgroup that 'permutes' with a larger group; the word 'normal' that is used today was introduced later by German mathematicians (we use it here in quotes).

Jordan analysed a chain of 'normal' subgroups, each one 'normal' in the one before, say

$$G = G_0 \rhd G_1 \rhd \cdots \rhd G_j \rhd G_{j+1} \rhd \cdots \rhd G_n = \{e\},$$

in terms of the orders of the quotients $\frac{|G_j|}{|G_{j+1}|}$, which he called the composition factors. In Book II, Section I, Subsection IV of his *Traité* he proved that when the chain is *maximal* – which means that it cannot be refined – the composition factors are unique up to order. It was only in 1873 that Jordan developed the theory in terms of the corresponding quotient groups, $\frac{G_j}{G_{j+1}}$, and regarded the composition factors as capturing the size of the quotient groups themselves.

Some examples may help. The composition series of S_4 is obtained from the maximal chain

$$S_4 \rhd A_4 \rhd V \rhd \mathbb{Z}_2 \rhd \{e\},$$

where V is the Klein group, and is therefore $2, 3, 2, 2$.

The composition series of S_5 is obtained from the maximal chain

$$S_5 \rhd A_5 \rhd \{e\},$$

because the group A_5 is simple, and is therefore $2, 60$.

To follow Jordan's argument that the quintic is not solvable by radicals, we need to introduce two more of his terms. He proposed to call an equation

abelian if its corresponding group is commutative – this is the origin of the term 'abelian' to refer to commutative groups. And he introduced the term *soluble groups* for the groups that characterise the equations that are solvable by radicals. His argument ran as follows.

See paragraphs 519–23 of Jordan (1870).

Theorem II For an equation to be solvable by radicals, it is necessary and sufficient that its resolution reduces to that of a sequence of abelian equations of prime degree.

Theorem III For an equation to be solvable by radicals, it is necessary and sufficient that its composition factors all be prime.

Corollary I Every group Γ contained in a soluble group G is itself solvable. For its composition factors, dividing those of G which are prime, are themselves prime.

Corollary II The general equation of degree n is not solvable by radicals if $n > 4$. For its composition factors, 2 and $\frac{n!}{2}$, are not all prime. (Jordan gave a 'proof' that the alternating group on n objects is simple when $n > 4$, but his argument is flawed, and seems to have been quietly corrected or replaced by subsequent authors.)

Corollary III The general equation of degree 3 is solvable by radicals. For its composition factors, 2 and 3, are prime.

Even before Jordan's book was published Galois's ideas had reached Germany, where they were taken up by Leopold Kronecker and Richard Dedekind, but it seems right to concentrate on Jordan's account, which was the most systematic.

1.3 Finite fields (Galois, Jordan, Steinitz)

The concept of a field has a curious history, as one might expect, because as we grow up we become very familiar with situations where we can add, subtract, multiply and divide (except by zero). The field of rational numbers does not seem to need a name or to require a new idea to become clear; it feels more like something that has always been there. Nor does much change when we learn about the fields of the real and then the complex numbers.

What slowly began to change mathematicians' opinions was the arrival of more unusual objects: the fields of algebraic numbers that arose as Galois Theory became better understood. These were explored for two reasons: the algebraic integers within a field of algebraic numbers had interesting number-theoretic properties, and rational functions in one or more variables were important in algebra and geometry. It seems that the first person to think there was now an abstract setting worthy of being given a name was Dedekind, who called a 'field' of algebraic numbers a 'Körper' in German. The word means a 'body', as does its translation 'corps' in French, and suggests an organic unity, which the English word 'field' does not – it is unclear how we came to settle on a blander term, but it was introduced by the North American mathematician Eliakim Hastings Moore in 1893.

It had also been noticed that there were some finite settings where the four basic operations of arithmetic could be carried out: the collection of integers modulo a prime. Gauss, in his *Disquisitiones Arithmeticae*, proved that the non-zero elements in such a collection are all powers of a single element, the so-called primitive root. But it was Galois, in one of the few papers he published in his lifetime, who saw more deeply.

Primitive roots are discussed in Book A, Chapter 4.

Galois claimed that if one took a polynomial f of degree ν that was irreducible over F_p, the field of p elements, and formed the collection of expressions of the form

$$a_0 + a_1 i + a_2 i^2 + \cdots a_{\nu-1} i^{\nu-1},$$

where the a_js are integers modulo p and i is a root of f, then these expressions behave like the natural numbers – by which he meant that one can add, subtract, multiply and divide with them (division by zero only being excluded). Moreover, every expression of this kind satisfies an equation of the form

$$x^{p^\nu} = x,$$

and there is at least one x of this form that satisfies no equation of the form $x^{p^n} = x, n < \nu$. This is the theorem of the primitive element: the multiplicative group of a finite field is cyclic.

See Neumann (2011, p. 65).

Later generations took this to mean that a finite field of p^ν elements can be constructed in this way that extends the field with p elements. The converse (every finite field has p^n elements and is generated by a single element of order p^n for some prime p and some integer n) was taken to be an open conjecture until it was established by Moore in 1893. It follows that up to isomorphism there is only one field of each possible order. The passage of over 60 years from Galois's death to Moore's paper is not an indication of the difficulty of the result, so much as a measure of what had to change in mathematics before fields – and finite fields in particular – came to stand out as worthy of attention.

Even by the end of the nineteenth century, fields (like rings) were still regarded as means to an end. A mathematician might have a problem in number theory or in polynomial algebra or geometry that seemed to require sophisticated handling, and might think it worth standing back from the details and trying a general argument about fields and field extensions of the following form: the basic collection of numbers or coefficients in the problem is the rational numbers; to this we adjoin some roots of polynomial equations, add some variables that themselves satisfy some equations, and so on. Textbook writers at the highest level (for example Heinrich Weber, a friend of Dedekind's) would then say there are settings in which we can add, subtract, multiply and divide (except by zero) in the ordinary way, and these are called 'fields' (other names, now abandoned, were also used). But apart from Galois's fields, they all have the rational numbers as their basic or ground field.

What catalysed a change in perception was a paper by the German mathematician Ernst Steinitz. His 'Algebraische Theorie der Körper' (The algebraic theory of fields) first appeared as an article in the major German

mathematics journal, Crelle's *Journal für Mathematik*, in 1910, and was eventually judged so important that it was reprinted as a book in 1930 with additional material by Reinhold Baer and Helmut Hasse, who hailed the paper as the starting point for many far-reaching researches in the domain of algebra and arithmetic. It marks the moment when fields themselves became objects of interest and were not merely means to other ends.

Fields are discussed in Book C, Chapter 11 and Book E, Chapters 18 and 19.

The paper opens with an introduction in which Steinitz says that his definition of a field is the same as Weber's: a system in which elements can be added, subtracted, multiplied and divided (only division by zero is excluded), in which the associative and commutative laws for addition and multiplication are obeyed, and in which multiplication is distributive over addition. He went beyond Weber in claiming that it was not Galois Theory, but the field concept itself that was now to be taken as central, so the aim of the paper was to survey all possible fields.

Steinitz began with the basic concepts: isomorphisms of fields, subfields, extension fields, adjunction of elements and integral domains. He then looked for the simplest possible fields, which he called *prime fields* and defined as fields that have no subfields. He proved that every field has a unique prime subfield, and that either there is a prime number $p > 0$ such that $px = 0$ for every x in the subfield, in which case the *characteristic* of the field is p, or the only number with this property is 0, in which case the characteristic of the field is 0.

Then Steinitz looked for extensions of the prime field. There are simple extensions, when a single element is adjoined, and these may be algebraic or transcendental. Algebraic extensions are finite-dimensional, and form the subject matter of Galois Theory. The rational functions in a variable x over a field K form a field $K(x)$, and if a polynomial $f(x)$ of degree n is irreducible over the field then $\dfrac{K(x)}{f(x)}$ is also a field; it is an algebraic extension of the ground field K of degree n. The new field is said to be obtained from the old one by adjoining all the roots of the polynomial $f(x)$.

Steinitz was at his most original and insightful when it came to dealing with the differing properties of fields of zero characteristic, and those of characteristic p. He gave a thorough account of fields of finite characteristic and rewrote the theory to include them systematically, with a discussion of the novel issues that arise when working with a field of characteristic p.

Amusingly, given the thrust of Steinitz's original presentation, when Baer and Hasse reissued the paper as a book they provided additional material on Galois Theory. They presented it in terms of a ground field K and a normal algebraic extension field L (a normal extension was defined by Steinitz as one with the property that an irreducible polynomial over K either remains irreducible over L or factors completely). In their hands, Galois Theory became the study of how fields intermediate between K and L relate to subgroups of the Galois group G of the extension, which is defined as the automorphisms of L that leave K pointwise fixed. Polynomials, and solvability by radicals, finally disappeared from the theory.

2 The origins of topology and metric spaces

2.1 The importance of Fourier series

Topology, and especially the study of metric spaces, grew out of analysis, and within analysis the dominant problem was the behaviour of sequences (of numbers or, more usefully, of functions) when going to a limit. Typically, one would have a sequence of functions that, formally at least, seemed to have a limit – for example, a sequence of polynomials $\sum_{j=0}^{n} a_j x^j$ with the power series $\sum_{j=0}^{\infty} a_j x^j$ as its limit as $n \to \infty$. Did the limit of this sequence of functions make sense – did it define a function? If so, what could be said about it?

Although it was known in the eighteenth century that the method of successively approximating a solution was often productive, the tools for bringing questions such as the one above into sufficient focus to be answerable were only developed in the early nineteenth century. By then there were many factors pushing mathematicians towards the epsilon-delta type of analysis that was finally introduced by Cauchy in the 1820s. Whenever functions were handled by their Taylor series approximations, or differential equations were solved by power series methods, the need arose to be sure that cutting off the series at a certain point was reliable. It was well known that some series make numerical sense only for a restricted range of the variable: for example

$$1 - x + x^2 - x^3 + \cdots ,$$

which equals $\frac{1}{1+x}$ if, and only if, $|x| < 1$. And, indeed, it was widely agreed that while the calculus of Newton and Leibniz was somehow right, there was no truly satisfactory explanation of why this was so.

Cauchy's work gave the basis of a new approach. He did not appeal to dynamical arguments in which points move along curves and reach limiting positions, and he did not appeal to geometrical arguments about tangents. Instead, he worked with an arithmetical approach based on inequalities, and did so much more systematically than anyone had done before. This had the virtue of making the calculus applicable to problems in mechanics and geometry without the risk of a vicious circle, and it steadily provided a set of procedures for proving what had hitherto been accepted largely on intuitive grounds. With Cauchy's analysis it became possible to prove the Intermediate Value Theorem, to state and prove the Mean Value Theorem, to prove that a function continuous on an interval is integrable there, to prove the fundamental theorem of the calculus, to establish numerous convergence tests for sequences and series; and so to resolve most of the problems that had begun to crop up whenever the calculus was used.

Fourier series

The next challenge came from one of the best ideas ever introduced into applied mathematics: Fourier series. Joseph Fourier argued for many years that the problem of the diffusion of heat through a solid could be solved by regarding the temperature as a function of time and position, expressing the flow of heat by means of a specific partial differential equation that he wrote down, and writing the solutions in the form of series involving the trigonometric functions. He eventually succeeding in publishing his ideas in his *Théorie analytique de la chaleur* (*The Analytical Theory of Heat*) in 1822.

In the simplest case, the solid body might be a slab heated at its base, so Fourier was interested in the temperature at a point as a function of time and the height above the base. He argued that the partial differential equation implies that the temperature is given by a possibly infinite sum of terms of the form $f(x)g(t)$, when $f(x) = \cos(mx)$ and $g(t) = e^{-mt}$, and so he looked at series of the form

$$ae^{-y}\cos x + be^{-3y}\cos 3x + ce^{-5y}\cos 5x + de^{-7y}\cos 7x + \text{etc.} \qquad (2.1)$$

Other Fourier series may involve the sine function exclusively, or both sines and cosines. This one involves only cosines because of other assumptions Fourier made about the shape of the slab.

subject to the boundary condition at the base that

$$1 = a\cos x + b\cos 3x + c\cos 5x + d\cos 7x + \text{etc.} \qquad (2.2)$$

The coefficients $a, b, c, d, \ldots$ have to be determined, and the series shown to converge. Fourier claimed that the coefficients were given in this case as

$$a = \tfrac{4}{\pi}, \quad b = \tfrac{-4}{\pi}\cdot\tfrac{1}{3}, \quad c = \tfrac{4}{\pi}\cdot\tfrac{1}{5}, \quad d = \tfrac{-4}{\pi}\cdot\tfrac{1}{7}$$

by the simple formula, but took it for granted that any series of this type and obtained in this way converged. He was dealing with practical problems where formal mathematics can be supported by empirical observations, but his work raised questions for others. Did the series of his type always converge? Could any function be written as a Fourier series? It is simply a statement of fact that questions raised in the study of Fourier series did more to drive mathematicians to take up rigorous analysis than any other development in the nineteenth century.

Abel was among the first to notice that something was amiss. Fourier had given the example of the series

$$\sum_{k}^{\infty} \frac{(-1)^{k+1}\sin(kx)}{k},$$

which we will call $Ab(x)$. The graph of this function is copies of the function $y = \frac{x}{2}$ between $-\pi$ and π and discontinuous at all odd multiples of π.

A Maple graph of the first 100 terms of this series shows how good the match is between the function and the Fourier series away from the points of discontinuity (Figure 2.1), but notice how unhappy Maple gets at the points where the function will be discontinuous (the vertical lines are a Maple artefact and should be ignored).

Maple is a very commonly used computer algebra package.

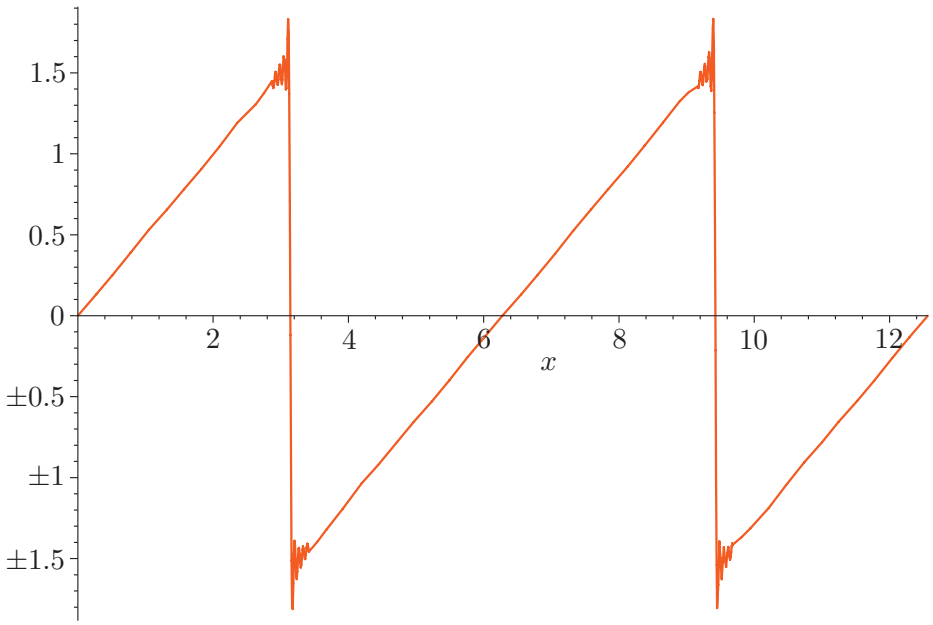

Figure 2.1 The first 100 terms of the series for $Ab(x)$

Since every term of the Fourier series is continuous (indeed, infinitely differentiable) and the series itself is visibly discontinuous at some points, we also have the interesting spectacle of a series of continuous functions not summing to a continuous function (contrary to what Cauchy thought he could prove in the 1820s).

The person who did most to start the productive investigation of these questions was Peter Gustav Lejeune Dirichlet. He wrote the Fourier series associated with a function $\phi(x)$ on the interval $[-\pi, \pi]$ as

$$\begin{aligned}&\tfrac{1}{2}b_0 + b_1\cos x + b_2\cos 2x + \cdots + b_m\cos mx + \cdots\\ &\quad + a_1\sin x + a_2\sin 2x + \cdots + a_m\sin mx + \cdots\end{aligned}$$

where the coefficients are given by Fourier's formulas:

$$b_m = \frac{1}{\pi}\int_{-\pi}^{\pi}\phi(\alpha)\cos(m\alpha)d\alpha, \tag{2.3}$$

$$a_m = \frac{1}{\pi}\int_{-\pi}^{\pi}\phi(\alpha)\sin(m\alpha)d\alpha. \tag{2.4}$$

He then proved, in a beautiful exercise in rigorous epsilon-delta analysis, that this series converges to the given function at every point where the function is continuous, provided that the interval $[-\pi, \pi]$ can be divided into a finite collection of sub-intervals on each of which the function is monotonic (Dirichlet's Theorem). He also established what happens if the function has a jump discontinuity at a finite number of points. Let $a \in [-\pi, \pi]$ be such a point, and suppose that

$$\lim_{x\uparrow a} f(x) = A_- \quad \text{and} \quad \lim_{x\downarrow a} f(x) = A_+,$$

where the limits are taken from below and above respectively. Dirichlet now showed that the Fourier series at $x = a$ takes the value $\frac{1}{2}(A_- + A_+)$.

This took care of the problem Abel had raised, but it opened many more. If the restriction to jump discontinuities is dropped, and a function such as

$$f(x) = \begin{cases} \sin\left(\frac{1}{x}\right), & \text{if } x \neq 0, \\ 0, & \text{if } x = 0, \end{cases}$$

which is discontinuous at $x = 0$, is admitted, then Dirichlet's Theorem has nothing to say. Functions with this kind of discontinuity are said to oscillate infinitely in the neighbourhood of the point where they are discontinuous, but there are also continuous functions that oscillate infinitely in an interval (for example, $x \sin\left(\frac{1}{x}\right)$) and they too are not covered by Dirichlet's Theorem. These functions have infinitely many maxima and minima. What can be said about them? Dirichlet himself pointed to an even worse example: Fourier's formulas for the coefficients only work if the function ϕ is integrable. What, he asked, can be said when ϕ is not integrable, and is, for example, the function that takes one value, c, when x is rational and another value, d, when x is irrational?

Uniform convergence

It is one thing to point towards work that has to be done, but quite another to find a way of doing that work. The next big advance in the study of Fourier series was made by Dirichlet's greatest student, the shy but brilliant Bernhard Riemann.

It is worth pausing here to note that at various stages in the mid-nineteenth century a great 'safety barrier' was created, which defined a class of functions with most of the intuitive properties one would expect. This barrier was formed from the concepts of uniform continuity and uniform convergence. We will define these terms for completeness, but the only point to note is that a 'fence' was erected within which functions can be proved to have a wide class of intuitively accessible properties. Attention then shifts to functions in the wild country beyond the fence.

A sequence of functions (f_n) defined on a set of points T converges *uniformly* to a function f if, for every $\varepsilon > 0$, there is an N such that $n > N$ implies $|f_n(x) - f(x)| < \varepsilon$ for every x in T. It soon became clear that if each of the functions f_n is continuous at a point t_0 in T and the f_n converge uniformly in T then the limit function f is also continuous at t_0. It was also found that the uniform convergence of a sequence of integrable functions f_n to a function f on a set T (given a suitably precise theory of the integral, which we are carefully not specifying) guaranteed not only that the limit function was integrable, but also that the sequence of integrals $\left(\int_T f_n(x)dx\right)$ converged to the integral $\int_T f(x)dx$, so term-by-term integration of a uniformly convergent series is a valid operation.

The situation for sequences of differentiable functions was a little more complicated, and we omit the details in favour of the moral of the story, which is that uniform convergence of well-behaved functions produces well-behaved limits. There is no such guarantee without uniformity. The result was that mathematicians were forced to look for uniformity if they wanted the strong intuitive results, and to work without it if they wanted to look for functions that were intelligible (because they were limits of

well-behaved functions) but not themselves well-behaved (and were perhaps continuous but not differentiable, or integrable but not continuous, or not even integrable). The second, much more arduous route was first taken by Riemann.

Non-uniform convergence and strange functions

Between 1851 and 1854 Riemann wrote his post-doctoral work under Dirichlet's direction. (This gave him the necessary qualification to teach in a German university.) He prefaced his final essay with a lengthy historical account of the history of Fourier series, going back to Euler. He then argued that although functions which do not satisfy Dirichlet's conditions are unlikely to occur in nature, they nonetheless have a 'very close connection with the principles of infinitesimal calculus, and can serve to bring greater clarity and rigour to these principles'. Moreover, he had found that such functions occur in parts of analytic number theory, so they were important in pure mathematics.

Riemann first showed how to define integrability in such a way that certain discontinuous functions could be integrable. Then he turned his attention to functions with infinitely many points of discontinuity, and gave a provocative example. He first defined the function $r(x)$ as follows. For each real number x,

$$r(x) = \begin{cases} x - m(x), & \text{if } x \neq \frac{n}{2}, \\ 0, & \text{if } x = \frac{n}{2}, \end{cases}$$

where $m(x)$ is the integer minimising the expression $|x - m(x)|$, and n is odd. So the function $r(x)$ takes values between $-\frac{1}{2}$ and $\frac{1}{2}$, and is zero where $x = \frac{n}{2}$ (where n is an odd integer). Its graph is illustrated in Figure 2.2 (again ignore the vertical lines).

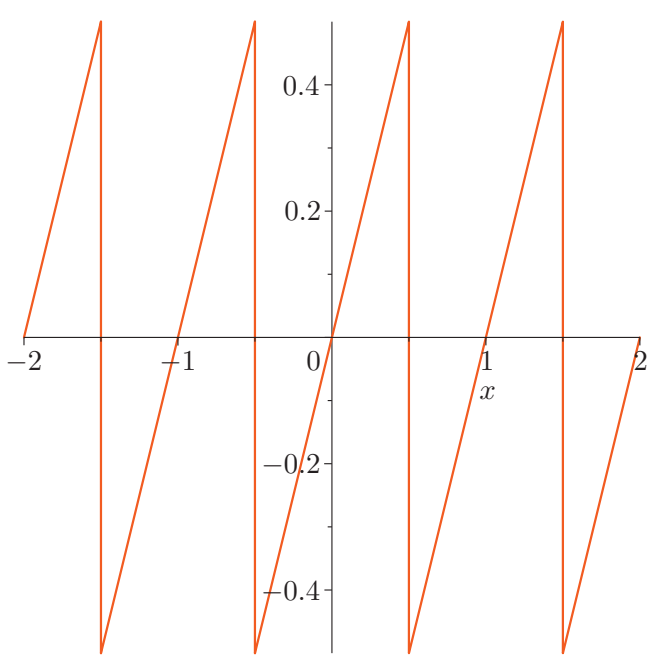

Figure 2.2 Riemann's function $r(x)$ on the interval $(-2, 2)$

Riemann then defined the function

$$f(x) = r(x) + \frac{r(2x)}{2^2} + \cdots + \frac{r(nx)}{n^2} + \cdots .$$

This function is discontinuous at every point $x = \frac{m}{2n}$, where m and $2n$ are relatively prime; this is a dense set of rational numbers. At these points, the left- and right-hand limits of the function f are

$$f(x_+) = f(x) - \frac{1}{2n^2}\left(\sum_k \frac{1}{(2k+1)^2}\right) = f(x) - \frac{\pi^2}{16n^2}$$

and

$$f(x_-) = f(x) + \frac{1}{2n^2}\left(\sum_k \frac{1}{(2k+1)^2}\right) = f(x) + \frac{\pi^2}{16n^2}.$$

The next Maple graph (Figure 2.3) shows the first 150 terms of the sum:

$$f(x) = r(x) + \frac{r(2x)}{2^2} + \cdots + \frac{r(150x)}{150^2}.$$

Again, ignore the vertical line that Maple has supplied where the function has a jump.

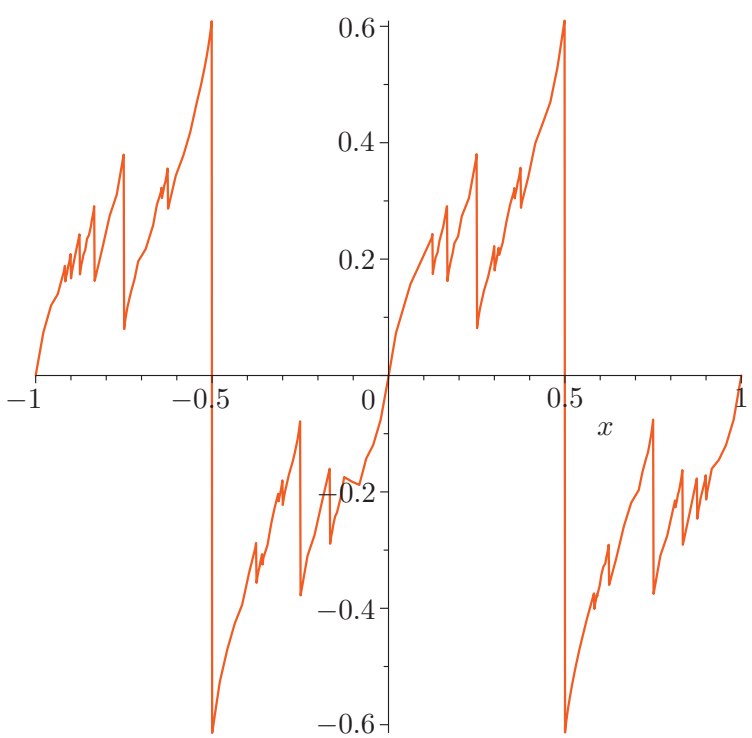

Figure 2.3 The first 150 terms of the sum defining Riemann's function

One can think of the integral $\int_a^x f(t)dt$ as x goes from a to, say, b as calculating a moving average. By the Intermediate Value Theorem, $\int_a^x f(t)dt = (x-a)g(x)$, where $g(x)$ is the average value of $f(x)$ on the interval (a, x). So a function should be integrable if it always has an average value at every point of an interval on which it is defined. Such

functions may oscillate wildly and be discontinuous, and even fail to be defined at finitely many points, so it is reasonable that integrable functions might be discontinuous at some points and oscillate to some degree. In the example just considered, each jump affects the average value by a finite amount, but the function is integrable because the contributions of the jumps are negligible except at finitely many points.

To study functions that lie beyond the domain of Dirichlet's Theorem it is necessary to have examples, and Riemann's breakthrough was his introduction of what are called trigonometric series, which he wrote this way:

$$\Omega(x) = A_0 + A_1(x) + A_2(x) + \cdots + A_n(x) + \cdots , \tag{2.5}$$

where

$$A_0 = \tfrac{a_0}{2}, A_n(x) = a_n \cos nx + b_n \sin nx.$$

These are not necessarily the Fourier series of any function. Riemann assumed that $A_n(x)$ converged to 0 as n increased – in fact, he treated this as uniform convergence throughout the paper. The convergence is for all values of x under consideration.

Very ingenious arguments allowed Riemann to derive necessary and sufficient conditions for a function to be represented at a point by a trigonometric series Ω. In this way he opened the door to all sorts of unexpected, and largely unknown, functions. A list of his results will indicate the range of behaviour of functions that he was now able to express.

1. Functions with infinitely many maxima and minima can be integrable without being representable by a Fourier series. For example,

   $$\frac{d(x^\nu \cos \frac{1}{x})}{dx}$$

   on $0 \le x \le 2\pi$, with $0 < \nu < \frac{1}{2}$.

2. Conversely, there are functions that cannot be integrated but can be represented by a trigonometric series. For example, the function

   $$\sum_{k=1}^{\infty} \frac{r(kx)}{k}$$

   exists for every rational value of x and is representable by

   $$\sum_{1}^{\infty} \frac{s(x)}{n\pi} \sin(2n\pi x),$$

   where $s(x)$: $= \sum \sigma^d - (-1)^d$ and is taken to be positive or negative according to whether n is odd or even, and σ ranges over all the divisors of n. This series is not bounded in any interval, no matter how small, and consequently is not integrable.

   Another example is

   $$\Sigma c_n \cos(n^2 x),$$

where $c_n \to 0$ but $\Sigma c_n = \infty$. As Riemann explained (1854, **S**13), 'For, if $\frac{x}{2\pi}$ is rational with denominator m in lowest terms, then clearly the series either converges or diverges to infinity according as

$$\sum_{n=0}^{m-1} \cos(n^2 x)\,, \quad \sum_{n=0}^{m-1} \sin(n^2 x)$$

are zero or not. Both cases arise.'

3. Finally, a trigonometric series can converge for infinitely many x in any interval and yet the coefficients not tend to zero. For example

$$\sum_{1}^{\infty} \sin(n!\pi x),$$

which converges not only for every rational value of x but also for some irrational numbers, such as $\sin 1, \cos 1, \frac{2}{e}$ and its multiples, etc.

Figure 2.4 is a graph of the first four terms of Riemann's series in result (3), which converges at some points even though the coefficients do not tend to zero.

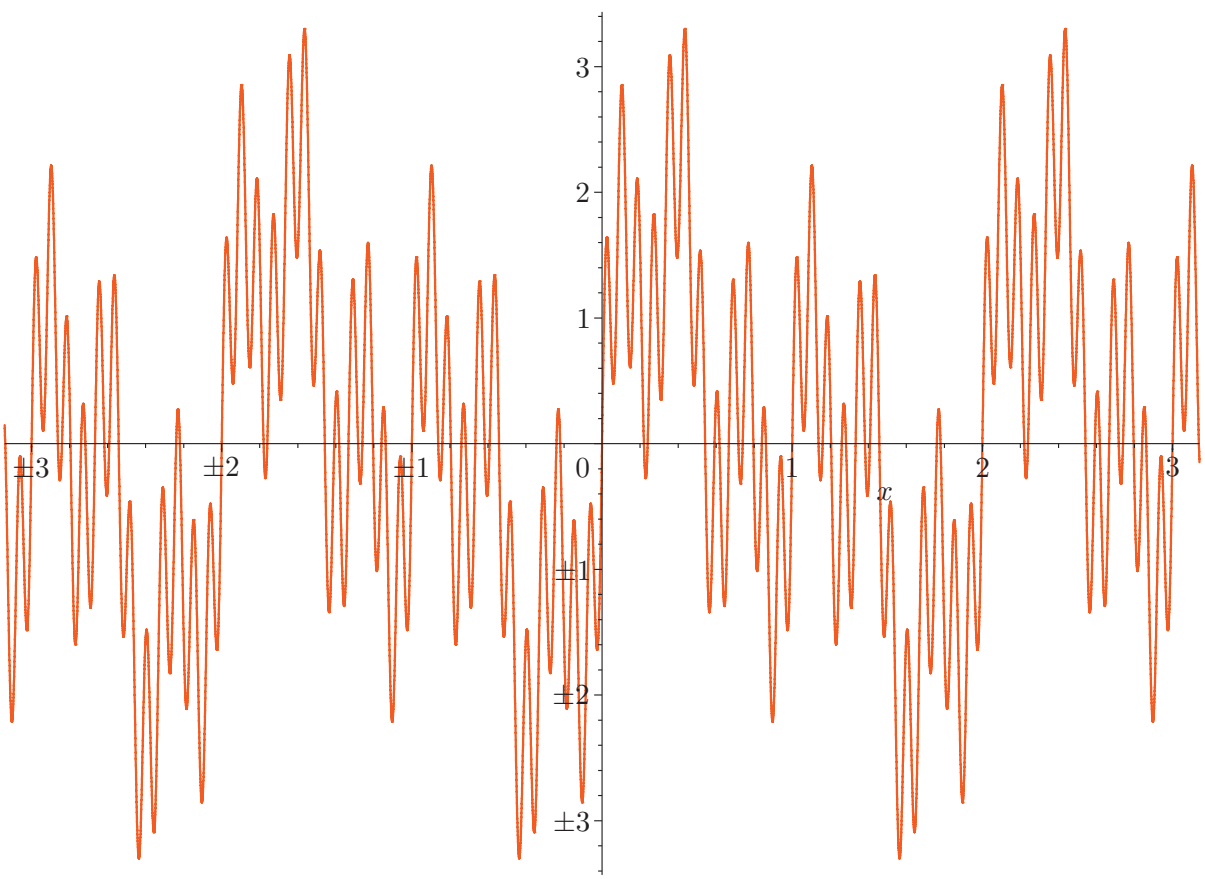

Figure 2.4 The first four terms of $\sum_{1}^{\infty} \sin(n!\pi x)$

Much of the ensuing work by a generation of analysts centred on exploring the highly counter-intuitive properties of functions that are continuous but

not differentiable. Riemann suspected, and in 1872 Weierstrass eventually proved, that there are functions continuous everywhere on an interval but differentiable nowhere. Weierstrass gave the example of the functions

$$f(x) = \Sigma_n b^n \cos(a^n \pi x) \ , \quad 0 < b < 1 \ , \ ab > 1 + \tfrac{3\pi}{2} \ , \ a \text{ an odd integer.}$$

Continuity follows from the fact that, by Weierstrass's M-test with $M_n = b^n$, the series is uniformly convergent for all x, but establishing non-differentiability everywhere is hard. The condition that the integer a be odd makes the fact that $\cos(N\pi) = \pm 1$ easier to handle. The condition that $ab > 1 + \frac{3\pi}{2}$ is what makes the tail of the series oscillate so much that the function is nowhere differentiable.

A comparison of the Maple graphs of the Weierstrass function for $a = 13, b = \frac{1}{2}$, where Weierstrass's conditions are obeyed, and for $a = 3, b = \frac{1}{2}$, where they are not, is instructive.

Figure 2.5, which shows the graph of just the first five terms of the series, powerfully suggests that differentiability is failing everywhere. By contrast, Figure 2.6, which shows the first 15 terms in the fake, suggests that differentiability is failing only at some points (albeit possibly a dense set).

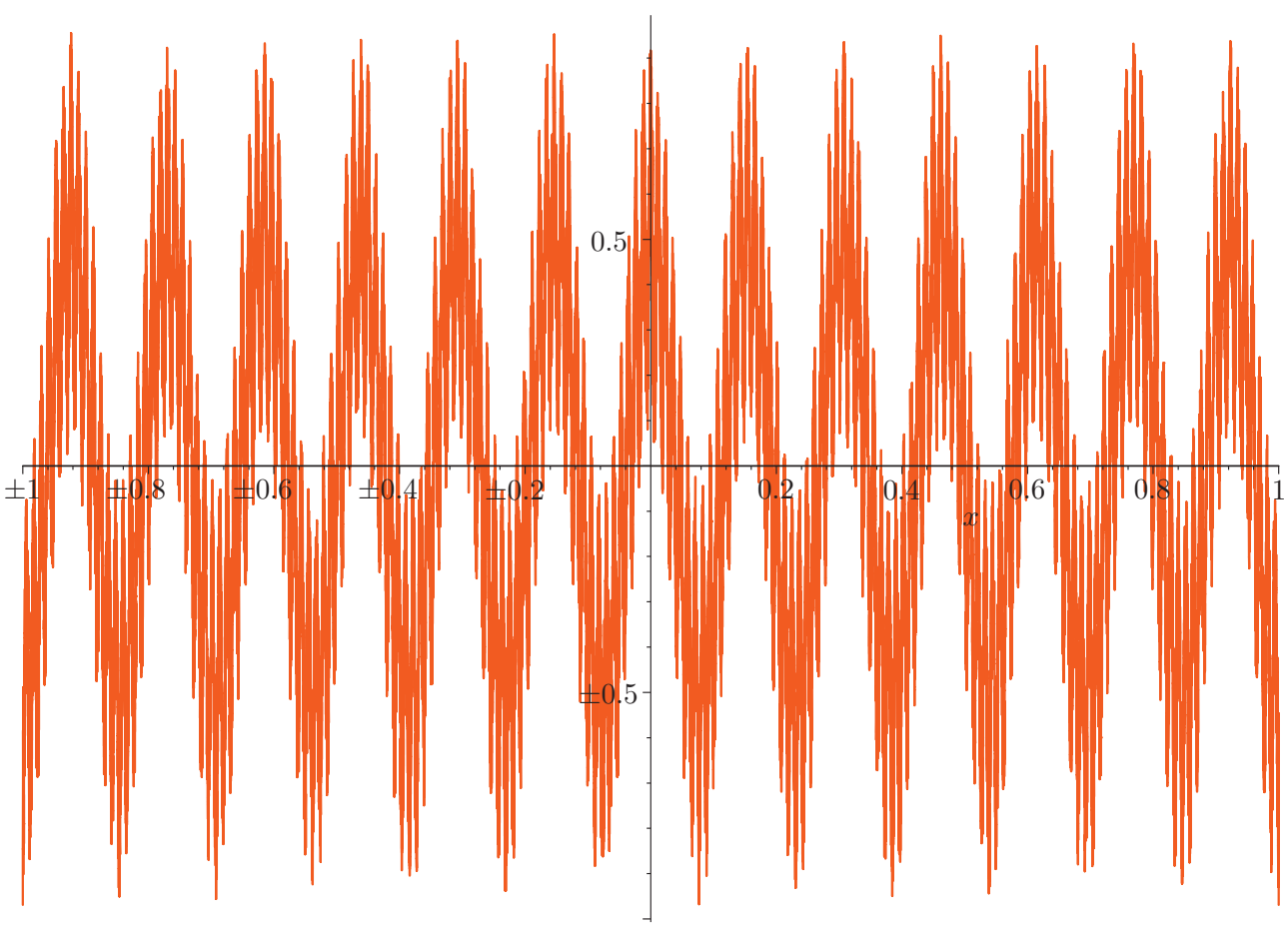

Figure 2.5 The first few terms of a Weierstrass function

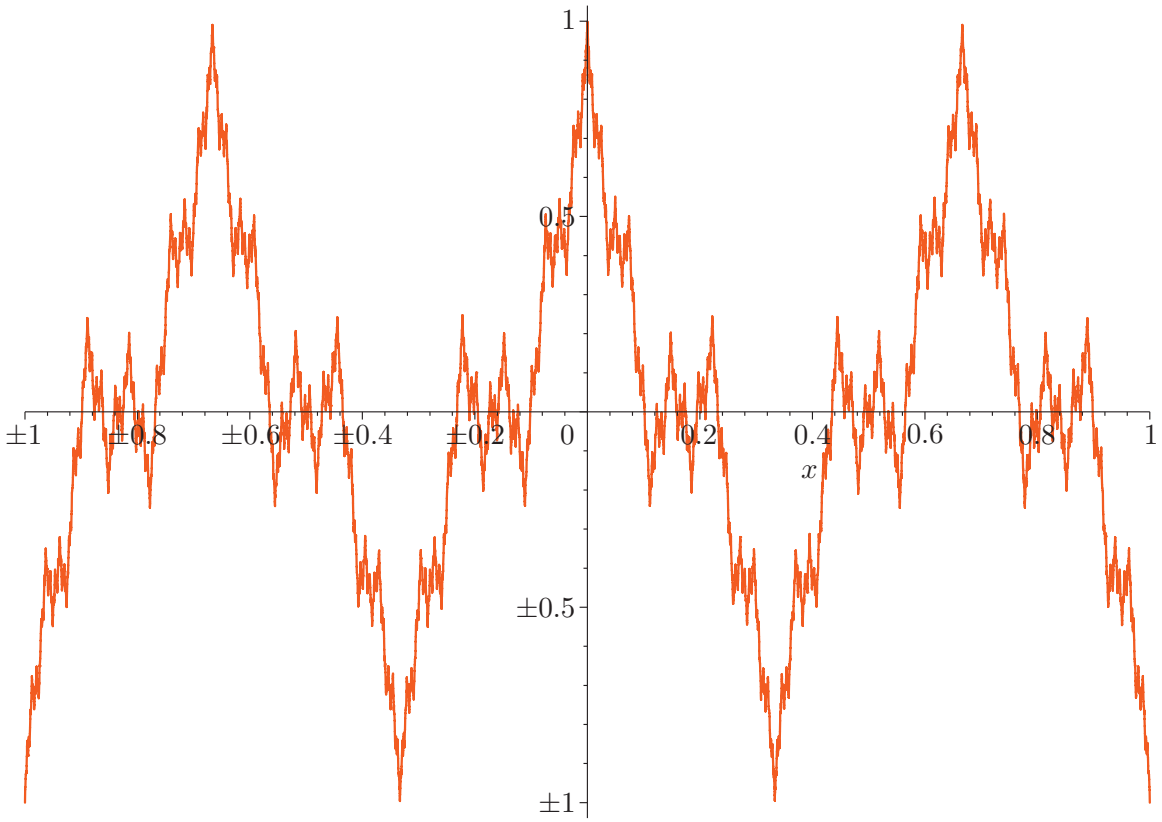

Figure 2.6 A 'fake' Weierstrass function

Cantor and point-set topology

Georg Cantor was the first to explore deeply the implications of the fact that if two integrable functions agree everywhere on an interval except at finitely many points then they have the same integral. His explorations led him to discover some of the fundamental features of point-set topology, and a decade later to his discovery of infinite sets of various sizes.

Cantor's work took off from the investigations of a colleague of his at the University of Halle in 1870. Heinrich Heine had raised the question of what can be said if the formulas for the Fourier coefficients (equations 2.3 and 2.4) are invalid, as they may be if the series is not uniformly convergent, and so term-by-term integration is illegitimate. He answered this question by giving some criteria about the set of points where continuity failed that, he thought, still allowed mathematicians to establish that the coefficients were unique. In 1872 Cantor aimed at showing uniqueness for any set of discontinuities. He investigated what can be said if a function is represented by two trigonometric series, $\sum a_j \sin jx + b_j \cos jx$ and $\sum a'_j \sin jx + b'_j \cos jx$, that converge and agree everywhere on an interval I except at a set P, where either convergence fails for at least one of them or the two trigonometric series take different values.

Cantor began with sets P that have only finitely many limit points. He called the set of limit points of a set P the derived set of P, and denoted it

P', so he began with sets whose derived set is finite. He called these sets of type 1, and proceeded inductively to define sets of higher type. A set of type 2 is one whose derived set is of type 1, and Cantor denoted the derived set of P' by $P^{(2)}$. A set of nth type is one for which its nth derived set $P^{(n)}$ is finite. At this stage Cantor had almost no examples of the sets he was talking about, and noted only that the set $\frac{1}{n} : n = 1, 2, \ldots$ is of type 1, and that the set of all the rational numbers in $[0, 1]$ is of no type. But he was able to show that the representation is unique if convergence holds except on a set of type n for some n. His proof was a modification of Heine's, extended by induction to any integer n.

By definition, a set of type 1 has a finite derived set, so there are only finitely many points p' each with the property that every neighbourhood of p' contains infinitely many points of P. These points p' need not belong to P, but the points of P somehow bunch up around them. It helps to introduce a term that Cantor was only going to introduce later: we say that a set is *nowhere dense* if its closure has an empty interior. Therefore, such a set contains no interval. It is clear that a set of type 1 is nowhere dense, and indeed that a set of type n is nowhere dense. So Cantor's theorem about the uniqueness of representation says very roughly that the representation is unique, provided that convergence holds everywhere except on a set with no interior.

Smith discovers Cantor sets

One fascinating thing to observe here is that what is at stake is what sets, in the domain of definition of a function, can be said to be negligible for the purposes of integration. However, the answer being proposed is topological. It seemed to several mathematicians in the 1870s that a set would be negligible for the purposes of integration if, and only if, it was nowhere dense. To modern eyes, this hints at a confusion, and the first to highlight it was the Oxford mathematician Henry Smith (Figure 2.7).

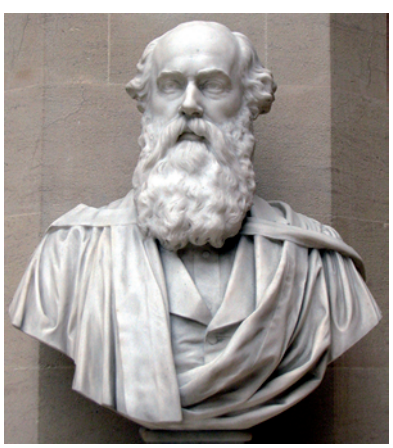

Figure 2.7 Henry John Stephen Smith (1826–1883)

Let us briefly introduce a term for a concept that was introduced in the 1870s without being given a name. If a set A is covered by a finite set $I = \{I_k\}$ of intervals, then the total length of the covering is $L_I = \sum_k |I_k|$. We are interested in the least upper bound of all finite coverings as the number of intervals in the covering increases arbitrarily because, if this quantity is zero, the set A will be negligible for the purposes of integration. Let us call such sets *zero coverable*.

Smith's first family of zero coverable sets was constructed in terms of a family of sets. The set P_1 consists of all the rational numbers of the form $\frac{1}{n_1}$, where n_1 is a integer; note that its points cluster around the point $x = 0$, which is not itself in P_1. The set P_2 consists of all the rational numbers of the form $\frac{1}{n_1} + \frac{1}{n_2}$, where n_1 and n_2 are positive integers. It is easy to check that P_1 is a subset of P_2, but there are points of P_2 that cluster around any element of P_1. The construction proceeds iteratively: the set P_k consists of all the rational numbers of the form $\frac{1}{n_1} + \frac{1}{n_2} + \cdots + \frac{1}{n_k}$, where $n_1, n_2, \ldots, n_k$ are positive integers.

You might like to confirm Smith's result that the jth derived set of $P_k, k > j$, is $P_k^{(j)} = \{0\} \cup P_1 \cup \cdots \cup P_{k-j}$. Smith showed that the sets P_k are zero coverable.

The second family of nowhere dense sets, Smith called sets 'in loose order' – see if you recognise them under another name. He defined them in these words (Smith, 1875, p. 147):

> Let m be any given [integer] greater than 2. Divide the interval from 0 to 1 into m equal parts; and exempt the last segment from any subsequent division. Divide each of the remaining $m - 1$ segments into m equal parts; and exempt the last segment of each from any subsequent division. If this operation be continued ad infinitum, we shall obtain an infinite number of points of division P upon the line from 0 to 1. These points lie in loose order ...

In more modern terms, Smith's process produces an infinite sequence of sets, each contained in the one before, and he took the set P to be all the points common to every set in the sequence. At each stage after the second, each set consists of a collection of intervals, and at each stage the new set differs from the one before because certain small intervals at the end of each interval have been removed.

The set P that Smith obtained is nowhere dense (informally, because intervals are being removed at each stage) and its complement at the nth stage is of length $(1 - \frac{1}{m})^n$, which tends to 1, so the set is zero coverable; a bounded function that is discontinuous at precisely the points of P is integrable.

Smith's set P is the first 'Cantor' set, S, after the very similar set that Cantor described in 1883. Cantor's construction resembles Smith's with $m = 3$, but at each stage the middle third of each interval is excluded. As this shows, Smith's works were not much read on Continental Europe.

Smith then changed the definitions to produce a new family of sets that are still nowhere dense but not zero coverable, so a function discontinuous on this set is not integrable. He let m be any integer greater than 2, divided the interval from 0 to 1 into m equal parts, and exempted the last segment from any subsequent division. Now he divided each of the remaining $m - 1$ segments into m^2 equal parts, and exempted the last segment of each from any subsequent division. So in this process, much smaller end intervals are being removed. The process is continued indefinitely, and the points that remain form an infinite set of points Q in the interval [0, 1]. This set Q is nowhere dense. But now at stage n the complement of the set is length $(1 - \frac{1}{m})(1 - \frac{1}{m^2}) \cdots (1 - \frac{1}{m^n})$, which does not tend to zero as n tends to infinity but instead tends to the positive limit $\Pi_n^\infty(1 - \frac{1}{m^n})$. For example, if $m = 3$ then this limit is over one-half (at about 0.560126). So the set Q is not zero coverable.

The example of a nowhere dense set in the unit interval that is not zero coverable shows that there was a confusion between a set being nowhere dense and its being zero coverable. Cantor's sets of type n are both nowhere dense and zero coverable, and this may have suggested to

mathematicians at the time that the sets which are negligible for the purposes of integration are the nowhere dense sets. What Smith showed was that this topological criterion was not the right one to use, and once this confusion was exposed the way was open for the topological concepts and the integration theory concepts to develop independently.

2.2 Potential theory and the Dirichlet principle

The topological concepts of connectedness and path-connectedness have their origins in another branch of mathematics: potential theory. The theories of gravitation and of electricity and magnetism in the nineteenth century had a common feature: it was assumed there was a force involved, which was represented by a vector. Because vectors are more complicated to handle than functions, George Green in England and Gauss in Germany argued independently in the 1830s and 1840s that the force was usually the derivative of a function. They claimed that the force $\mathbf{F}$ of, say, gravity due to a solid body could be written as $\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right)$ for a suitable function f. Green gave this function a name that it still retains: the potential function. For gravity and electricity, the potential function is harmonic:

$$\triangle f = \nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = 0.$$

Very quickly mathematicians became interested in proving the existence of such a function under very general conditions, corresponding to different distributions of mass or electric charge. Physicists, of course, were happy to assume potential functions always existed. The problem is different in two and three dimensions, and the simpler but very productive two-dimensional problem became this: given a domain D in the plane bounded by a closed curve C, and given a continuous function f on C, find a function φ defined on D that is harmonic on the interior of D and is equal to f on C. The idea is that the function φ represents some distribution of mass or electricity on C, and the function f is to be the corresponding potential function.

The problem proved very difficult to solve, despite the physicists' confidence. Green had claimed in 1833 that the three-dimensional problem can always be solved, and his argument was that the required function is the one that minimises the integral

$$\int_{vol} \left(\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 + \left(\frac{\partial f}{\partial z}\right)^2 \right) dx\, dy\, dz.$$

William Thomson (later Lord Kelvin) reaffirmed the principle himself in 1847, as did Dirichlet during lectures in 1856–57 at Göttingen (which were published posthumously in 1876).

In Dirichlet's opinion, the sought-for function u could seldom be found explicitly: 'we can only speak of an existence proof for it. The latter presents no difficulty.'

He went on (Dirichlet, 1876, pp. 127–8):

> For every bounded connected domain T there are clearly infinitely many functions u continuous together with their first-order derivatives, for x, y, z which reduce to a given value on this surface. Among these functions there will be at least one which reduces the following integral $U = \int_{vol}[(\frac{\partial u}{\partial x})^2 + (\frac{\partial u}{\partial y})^2 + (\frac{\partial u}{\partial z})^2]$ extended over the domain T, to a minimum; it is evident that this integral has a minimum since it cannot become negative. We can now show the following:
>
> 1. Every such function u which minimises U, satisfies the differential equation $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0$ everywhere in the domain T. This already makes it clear that there always exists a function u having the desired property, namely that function for which U becomes a minimum.
> 2. Every function u which satisfies the [above] differential equation within the domain T, minimises the integral U.
> 3. The integral U can have only one minimum.
>
> It follows from 2 and 3 that there is only one function u with the desired property.

To a physicist this 'principle' seems entirely reasonable. It confirms what nature herself tells you: there is an equilibrium distribution of charge, and so forth. But a mathematician needs to check whether there are any functions at all that extend the function defined on the surface to the whole of the interior or the exterior of the surface. Then it needs to be checked that there are such functions for which the integral is finite; if not, the class of functions Dirichlet wished us to consider is empty and the principle is vacuous. Last, but indeed not least, although Dirichlet said 'it is evident that this integral has a minimum since it cannot become negative', the examples of the function $f(x) = x$ defined on the open interval $(0, 1)$, and of $F(x) = \tanh(x) + 1$ defined on the whole of $\mathbb{R}$, show that a function may be never negative but still fail to have a minimum.

In 1857 Riemann attempted to prove the Dirichlet principle. (He did not simply assume it, as many books say.) He first asserted that if a certain integral over a surface T,

$$L(\alpha, \beta) = \int \left(\left(\frac{\partial \alpha}{\partial x} - \frac{\partial \beta}{\partial y} \right)^2 + \left(\frac{\partial \alpha}{\partial y} + \frac{\partial \beta}{\partial x} \right)^2 \right) dT, \tag{2.6}$$

is finite, where α and β are two arbitrary functions of x and y, then by varying α by a continuous function (or one that is discontinuous only at single points) that is zero everywhere on the boundary of T, the integral attains a minimal value – and moreover, this minimum is attained by a unique function if one excludes the points of discontinuity. He then showed easily that this unique minimising function is harmonic.

Riemann argued that the integral L was defined on a space of functions that was 'a connected domain closed in itself, in which each function can be transformed continuously into every other' (1851, **S**16), from which it followed that there was a sequence of functions in this space for which the

corresponding integral tended to a minimum. Because the space was connected and closed, the limit function was in this space, and so the principle was true.

A charitable interpretation of this remark would be that Riemann expected some theorems to be true because spaces of functions had properties like closed intervals. This is a remarkable insight – but it must also be said that he gave no reason to believe that the space of functions he was considering had the required properties. He did not even define what he meant by the phrase 'closed in itself'.

In any case, this attractive hint was destroyed by one of Riemann's students, Friedrich Prym. He exploited the behaviour of continuous, infinitely oscillating functions to define a continuous function on the boundary of the unit disc that does extend to a harmonic function on the interior of the disc but for which the Dirichlet integral is infinite. This shows that Dirichlet's principle can fail, even when Dirichlet's problem can be solved.

The next person to tackle the Dirichlet problem was Hermann Amandus Schwarz, who developed his solution to this problem in a series of papers around 1870. He first showed that the problem can be solved when the domain T is a disc. He let $f(\phi)$ be an everywhere finite, continuous, real-valued function that is periodic with period 2π – so he could think of f as defined on the unit circle, which is the boundary of the unit disc. He then defined a function $u(r, \phi)$ on the unit disc by the equations

$$u(r,\phi) = \frac{1}{2\pi}\int_0^{2\pi} f(\psi)\frac{1-r^2}{1-2r\cos(\psi-\phi)+r^2}d\psi\,, \quad 0 \le r < 1$$

and

$$u(1,\phi) = f(\phi).$$

It is not entirely obvious that this function is finite and continuous on the whole closed unit disc $0 \le r \le 1$, but Schwarz showed that it is by a careful analysis of the situation as $r \to 1$. Straightforward differentiation then shows that it satisfies the differential equation $\Delta u = 0$ in the interior of the disc, and so the function $f(\phi)$ is a solution of the Dirichlet problem.

It followed from some earlier work of Schwarz's that the problem is therefore also solved for a square. Schwarz therefore proposed to show how the Dirichlet problem could be solved for domains that are composed of an overlapping family of domains, on each of which the problem has already been solved. This is his famous alternating method, which we now outline.

See Schwarz (1870).

Schwarz's alternating method

Schwarz now supposed he had two domains, T_1 and T_2, on each of which separately he had a solution to the Dirichlet problem for any prescribed functions g_1 on the boundary of T_1 and g_2 on the boundary of T_2 (Figure 2.8).

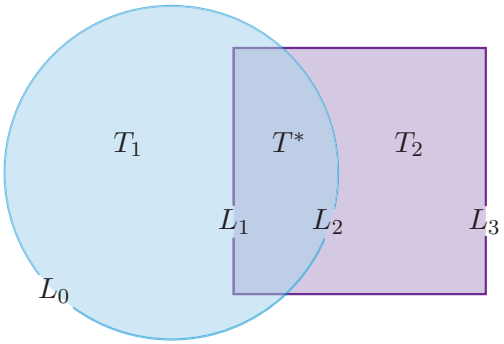

Figure 2.8 A circle and a square

He now set out an ingenious convergence argument. To make it work, he made some simplifying assumptions. First, the regions must overlap nicely; so $T^* = T_1 \cap T_2$, the region common to both of them, is two-dimensional and the boundaries of the regions cross with distinct tangents. Schwarz broke the boundary of each region into two parts: the part of the boundary of T_1 that lies outside T_2 he called L_0; the part of the boundary of T_1 that lies inside T_2 he called L_2; the part of the boundary of T_2 that lies outside T_1 he called L_3; the part of the boundary of T_2 that lies inside T_1 he called L_1.

His aim was to produce a harmonic function on the interior of $T_1 \cup T_2$ that has a continuous extension to a given function g on $L_0 \cup L_3$. Let the minimum value of g on $L_0 \cup L_3$ be k, which without loss of generality may be assumed to be positive.

Schwarz imagined that an air pump pumped air from the region T^* alternately into the two halves of T, the regions $T_1 \setminus T^*$ and $T_2 \setminus T^*$, through the membranes L_1 and L_2, until equilibrium was reached. More formally, he argued as follows.

1. Look at T_1. There is a harmonic function in T_1 that agrees with g on L_0 and takes the constant value k on L_2; call it h_1.
2. Now look at T_2. There is a harmonic function in T_1 that agrees with g on L_3 and with h_1 on L_1; call it h_2. Notice that the values of h_2 on L_1 are greater than or equal to k.
3. Now look again at T_1. There is a harmonic function in T_1 that agrees with g on L_0 and with h_2 on L_2; call it h_3.
4. Now look again at T_2. There is a harmonic function in T_2 that agrees with g on L_3 and with h_3 on L_1; call it h_4.

To conclude the proof, Schwarz showed that the series

$$h' = h_1 + (h_3 - h_1) + (h_5 - h_3) + \cdots + (h_{2i+1} - h_{2i-1}) + \cdots,$$
$$h'' = h_2 + (h_4 - h_2) + (h_6 - h_4) + \cdots + (h_{2i+2} - h_{2i}) + \cdots,$$

converge unconditionally and indeed define a harmonic function on the whole of $T_1 \cup T_2$ that has the prescribed boundary values of the function g on the boundary $L_0 \cup L_3$. So the Dirichlet problem is solved for the union of the two regions, which is what Schwarz set out to do.

Schwarz's proof was a significant step forward, but it remained unclear what class of domains could be treated in this fashion. In particular, could the Dirichlet problem be solved for a region obtained from the overlap of infinitely many regions for which it was solvable? This was the criticism made by Axel Harnack in his book of 1887.

Connectedness – Harnack's work

Harnack considered a sequence of harmonic functions u_n defined on a region F of the plane, where each function u_n restricts to a continuous function U_n on the boundary of F. His idea was to show that as the sum of the functions on the boundary converges to a given function U, the corresponding functions in the interior converge to a harmonic function. He first proved, under suitable conditions on the behaviour of the functions u_n near the boundary, that if the sum $U := \sum_n U_n$ converges uniformly then the sum $u := \sum_n u_n$ converges at every interior point of the surface F to a harmonic function. From this, what became known as Harnack's Theorem followed.

Theorem 2.1 *Harnack's Theorem*

If the members of a sequence of harmonic functions u_n all have the same sign (say, positive) and the sum $u := \Sigma u_n$ converges at an interior point of the surface F, then it converges at every interior point of the surface F to a harmonic function. Alternatively, if the harmonic functions u_n tend from below to the values of a function u, then u is a harmonic function.

Harnack tried to consider regions of a very general kind. He was the first mathematician to try to explain what it is for a domain to be connected. Under his definition, a space is *connected* if any two points can be joined by a finite polygonal arc that can be covered by overlapping discs that all lie in the interior of the domain. We would say that this defines a path-connected space. Later authors were to tease out the concept of a domain as a space in which every point has a disc-like neighbourhood lying entirely in the space.

Harnack defined a *boundary point* of a space as a point such that every neighbourhood of it contains some points that are in the domain and some that are not. He then claimed that the boundary of a simply connected domain is continuous, although it may be nowhere differentiable and may have corners and cusps, and could even have an arbitrary number of incisions into the domain, which would be lines drawn inwards from boundary points that would be traversed twice by any circuit of the boundary.

With these definitions in place, Harnack then proved that the Dirichlet problem can be solved for an arbitrary bounded region of the plane with an arbitrary boundary. He supposed that the general simply connected domain can be steadily approximated by polygonal regions, and used a

limiting argument, which he attributed ultimately to Schwarz, to deal with the general region. He used his theorem to establish that the sequence of harmonic functions on the polygonal regions converges to a harmonic function on the given domain. Finally, he patched together arbitrary bounded domains out of simply connected pieces.

Discs and their boundaries

One consequence of Harnack's work was that it became clear to mathematicians that there is a great variety of curves that can be the boundary of a disc. There is a large class of such boundaries that are in a continuous one-to-one correspondence with a circle. And then there are domains with one or more spikes, such as Figure 2.9 indicates.

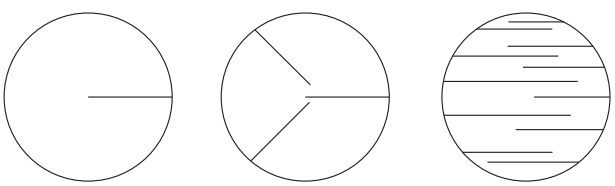

Figure 2.9 Some topological discs

Each of the examples in Figure 2.9 displays a region that is homeomorphic to a disc, but the boundaries are not simple closed curves. One can even contemplate regions with infinitely many spikes. Harnack's work was extended to this setting by the North American mathematician William Fogg Osgood in 1900.

Conversely, one can ask if a region bounded by a simple closed curve is necessarily homeomorphic to a disc. The first person to see that there was something to prove here and to attempt a proof was Camille Jordan, in a note appended to Volume 3 of his *Cours d'analyse* (1887, pp. 587–94). There he outlined a proof, which began by passing from the curve to an arbitrarily close polygonal approximation. Over the years this proof fell into suspicion, and one given by the North American mathematician Oswald Veblen in 1905 came to be regarded as the first correct argument, but a careful recent analysis by Thomas Hales (2007) has shown that Jordan's proof was fundamentally sound and can be polished up into a rigorous argument.

Compactness

The concept of compactness, on the other hand, has a curious history. There are several theorems which say that if a subset X of $\mathbb{R}$ or $\mathbb{R}^n$ is compact then a continuous function on X is uniformly continuous, and is bounded and attains its bounds. For example, Dirichlet in the 1830s proved this claim about continuous functions on a closed interval. The key ingredient in the proof is that any covering of a closed interval by open intervals contains a sub-covering that has only finitely many open intervals. The Heine–Borel Theorem says that a subset of $\mathbb{R}^n$ is compact if, and only if, it is closed and bounded.

The great advantage of compactness is that, at least intuitively, the limit of a sequence of points in a compact set is also in the set. Strictly speaking, the relevant concept here is that of sequential compactness: a space is sequentially compact if every sequence of points in it has a convergent subsequence. Happily for us, if a space is a metric space then it is compact if, and only if, it is sequentially compact.

The advantages of doing analysis on compact sets and compact spaces is clear. Approximations that tend to a limit tend to a limit in that set or space. Moreover, the Bolzano–Weierstrass Theorem says that every sequence on a compact set has a convergent subsequence. Problems in analysis where the solution is found by a sequence of approximations therefore differ markedly depending on whether they are on compact or non-compact spaces, and the sad but predictable news is that analysis mostly happens on non-compact spaces.

Consider, for example, the Dirichlet problem, where we are asked to find a function that minimises a certain integral over a given region. Candidate functions are required to be continuous everywhere inside the region, and agree with a given continuous function on the boundary of the region. One way of thinking of Harnack's Theorem in the present setting is that it tells us that the limit function does not always exist, and that is because the space of candidate functions is not compact.

2.3 Metric spaces – Fréchet

The mathematician who saw first how important compact spaces can be was the French mathematician Maurice Fréchet (Figure 2.10). Indeed, the idea of a metric space was born in his doctoral dissertation, where he discussed the behaviour of limits in what he called 'abstract sets' – he did not speak of spaces or topology until much later.

Figure 2.10 Maurice Fréchet (1878–1973)

In a 1906 paper, Fréchet first defined axiomatically abstract spaces with a concept of limit; then, finding that not much could be said at such an abstract level, he brought in the idea of *separation* ('écart') by adding an axiom that allowed him to talk of the separation between any two elements of an abstract set. He denoted the separation of two elements A and B by (A, B) and imposed the following rules:

$$(A, B) \geq 0,\ (A, B) = 0 \text{ if, } \textit{and only if, } A = B,$$

and if (A, B) and (B, C) are less than some ε then (B, C) is less than some $f(\varepsilon)$, where (ε) is a function that tends to zero as ε tends to zero. He then said that a sequence of elements A_j tends to a limit A if, and only if, $\lim_{j\to\infty}(A_j, A) = 0$. Spaces with a concept of separation Fréchet called *neighbourhood sets* or V-sets (the V standing for 'voisinage' or neighbourhood in French).

At one point in his doctoral thesis, Fréchet introduced the idea of a V-set that obeyed the triangle inequality:

$$(A, B) + (B, C) \geq (A, C).$$

Such a space is a metric space.

Fréchet then developed a number of theorems that are fundamental to point-set topology. For example, he defined a set to be *closed* if every limit point of the set belongs to the set, and *compact* if the intersubsection of every nested sequence of closed sets contains at least one point, and then proved, using methods introduced by Weierstrass, that a continuous function on a compact L-space is bounded and attains its bounds (L-spaces are spaces in which limits exist).

It is perhaps more interesting to note some of the examples Fréchet gave of spaces with limits. These included spaces of functions, spaces of curves or surfaces, and spaces of parameterisations of a given curve, where the notion of convergence is defined by a suitable definition of separation. We illustrate just one of these: the space of all curves in $\mathbb{R}^3$. A curve is defined as the image of three functions $(f(t), g(t), h(t))$ – that is, by a parameterisation, which is moreover defined on the closed unit interval. The separation of two curves C_1 and C_2 with parameterisations $(f_1(t), g_1(t), h_1(t))$ and $(f_2(t), g_2(t), h_2(t))$ is defined as the maximal value of

$$|f_1(t) - f_2(t)| + |g_1(t) - g_2(t)| + |h_1(t) - h_2(t)|,$$

which is finite because the curves lie in a finite region of $\mathbb{R}^3$, and the separation of the curves C_1 and C_2 is defined as the maximum value of their separations under all possible parameterisations.

2.4 Topological spaces – Hausdorff

Felix Hausdorff's *Grundzüge der Mengenlehre* (*Foundations of Set Theory*) of 1914, which he dedicated to Cantor, deals, as its title indicates, extensively with the novel theory of sets before turning to more topological questions in the second half of the book. Hausdorff thought that the greatest success of set theory so far was in clarifying geometrical ideas – where, in his opinion, the new methods were both more profound and simpler than the ones they replaced.

With no more than a mention of Fréchet's name in a footnote, Hausdorff gave a definition of a metric space ('metrische Raum'; the modern term may derive from this work) as one where to each pair of points x, y there is a non-negative number $\overline{xy}$ called their separation or distance ('Entfernung') such that:

- $\overline{xy} = \overline{yx}$
- $\overline{xy} = 0$ if, and only if, $x = y$
- $\overline{xy} + \overline{yz} \geq \overline{xz}$.

He noted that Euclidean n-dimensional space with the usual distance function is a metric space.

Hausdorff then pointed out that in a metric space one can define a neighbourhood of a point x as the set $U_x = \{y \,|\, \overline{xy} < \rho\}$, where ρ is a number greater than zero.

He then defined a topological space as a set with a family of subsets U_x, called the neighbourhoods of x, with the following properties.

1. There is at least one neighbourhood U_x for each x in the space, and every neighbourhood U_x contains x.
2. If U_x and V_x are two neighbourhoods of the same point x then there is a neighbourhood of x contained in $U_x \cap V_x$.
3. If y is a point in U_x then there is a neighbourhood of y that lies entirely inside U_x.
4. Given any two distinct points x and y, there are two neighbourhoods U_x, U_y such that $U_x \cap U_y = \emptyset$.

The last of these axioms now carries Hausdorff's name – it is the Hausdorff Axiom. As he noted, metric spaces are spaces of this kind; later mathematicians dropped this requirement from the definition of a topological space, and in modern mathematics we do not require a topological space to satisfy the Hausdorff Axiom.

Hausdorff now defined many useful topological terms, including interior points of a set, boundary points, closed sets, perfect sets and connected sets. A non-empty set is connected if it cannot be expressed as the union of two disjoint non-empty closed subsets. To define a compact set, he said he followed Fréchet but then gave a different definition: a set in which every infinite subset has an accumulation point (the accumulation point need not be in the set itself). The empty set and all finite sets were also declared to be compact. Under this definition, as Hausdorff noted, every subset of a compact set is compact, and he went on to prove that in Euclidean space (which is not compact even under this definition) a bounded set is compact.

Later in the book he looked at compact metric spaces. He defined a metric space to be bounded if it is contained in a spherical neighbourhood of one of its points, and totally bounded if for every positive number ρ the space is contained in finitely many spherical neighbourhoods of radius ρ. Here Hausdorff observed that a bounded subset of a Euclidean space is totally bounded. He then proved that every compact metric space is totally bounded. Under his definitions, the converse is false. As he noted, the set of rational numbers is totally bounded but not compact.

Our definition of a compact subset of Euclidean space, or any metric space, leads to the theorem that such a subset is compact if, and only if, it is closed and bounded. This leads us to ask when Hausdorff's concept of 'compact' acquired the extra concept of being closed, and so became Fréchet's concept, but a clear historical account seems not to have been written.

In the last thirty years, the study of fractals has made Hausdorff's innovative study of the dimensions of sets into his most cited paper. He published this paper (the title translates as 'Dimension and exterior measure') in 1919. In it, he looked at Euclidean spaces $\mathbb{R}^n$. He supposed that there was a family $\mathcal{U}$ of subsets with the property that for every $\varepsilon > 0$, every subset A of $\mathbb{R}^n$ can be covered by a finite or countable set of sets U

in $\mathcal{U}$, where each U has a diameter of less than ε. He also assigned to each U in $\mathcal{U}$ a non-negative real number $\ell(U)$. Then he defined, for A a subset of $\mathbb{R}^n$ and $\varepsilon > 0$,

$$L_\varepsilon(A) = \inf\{\sum_{j\geq 1} \ell(U_j)\},$$

where the sum is taken over a covering of A by sets in $\mathcal{U}$ of diameter less than ε. He then defined

$$L(A) = \lim_{\varepsilon \to 0} L_\varepsilon(A).$$

This, as Hausdorff showed, defines what is called a positive outer measure on subsets of $\mathbb{R}^n$; that is to say, the following four properties hold:

1. $0 \leq L(A) \leq \infty$, and $L(\emptyset) = 0$
2. $B \subset A$ implies $L(B) \leq L(A)$
3. $A = \bigcup_{j=1}^{\infty} A_j$ implies $L(A) \leq \sum_{j=1}^{\infty} L(A_j)$
4. if sets A and B are a non-zero distance apart then $L(A \cup B) = L(A) \cup L(B)$

along with one too technical to state here, which implied that under certain conditions $L(A)$ was the infimum of all values of $L(B)$ where $B \supset A$.

Then Hausdorff introduced what has become called a Hausdorff function, λ, which is a strictly increasing function from $[0, \infty]$ to $[0, \infty]$ with $\lambda(0) = 0$, and supposed that $\ell(U)$ was of the form $\lambda(d(U))$. This gave him outer measures L^λ, and he said that a subset A of $\mathbb{R}^n$ was of dimension λ if $0 < L^\lambda(A) < \infty$.

Hausdorff was not able to show that there are subsets of $\mathbb{R}^n$ of arbitrary dimension for any choice of Hausdorff function, but he was able to do so if the Hausdorff function was unbounded. His construction was a variant of the construction of the Cantor set. He defined a set of strictly positive numbers $a_j, j = 0, 1, \ldots$ such that $a_j > 2a_{j+1}, j = 0, 1, \ldots$ and $2^j \lambda(a_j) = 1, j = 0, 1, \ldots$. Now he took the interval $[0, a_0]$ and removed the central open interval of length $a_0 - 2a_1$, leaving two closed intervals of length a_1. From each of the remaining two intervals he removed the central intervals of length $a_1 - 2a_2$, leaving four closed intervals of length a_2, and so on. If the union of the 2^j intervals at the end of stage j is called A_j, then Hausdorff considered the set

$$A = \bigcap_{j=1}^{\infty} A_j$$

and showed that it has Hausdorff dimension λ.

It took almost a decade for mathematicians, starting with Besicovitch, to take up these ideas, but their importance has grown ever since. Curiously, this was Hausdorff's only paper on the idea, and he turned instead to work on rather more classical problems in analysis.

3 Biographies

3.1 Pythagoras

One of the earliest, and certainly one of the most famous, mathematicians is said to be Pythagoras, but we know very little about him. It is likely he wrote nothing; certainly no written work ascribed to him survives, and most of the early historical accounts that refer to him make him a religious leader who flourished in Croton in what is now South East Italy around 560 BCE. His followers, the Pythagoreans, adhered to his teaching for some centuries after his death; Aristotle wrote about them some 150 years later, but never directly about Pythagoras himself in any text we have.

The Pythagorean community after his death did contain a number of mathematicians. However, there is actually no evidence that Pythagoras knew, or proved, the theorem that today bears his name, and the famous stories about what happened when the result was discovered, and later revealed, are just that: stories. They say much more about the storytellers, and the needs of their audiences, than they do about the past. In particular, there is no truth at all in the dramatic story that Pythagoras taught that 'All is number', meaning that the discovery of irrational numbers such as the square root of 2 imperilled his entire philosophy and someone was drowned at sea for revealing this awful secret.

The Pythagoreans as a cult continued for many centuries, doubtless with many changes on the way, and there is good evidence that they were absorbed in number mysticism. According to Aristotle, the female was likened by them to the number two, the male to the number three, and their sum, five, was likened to marriage.

What we do know about ancient mathematics is that there is an abundance of evidence that Mesopotamian mathematicians a millennium before had known the Pythagorean theorem as a fact about right-angled triangles, as did mathematicians in other cultures. It is less clear that these cultures had a proof of the result. Somewhere along the line, perhaps in a Greek community, someone must have discovered a proof that $\sqrt{2}$ is irrational, because Aristotle sketched a proof, but the Greeks handled irrational square roots by the simple device of comparing their squares; there is no evidence that they went into a panic about some failure of geometry to reduce to the study of whole numbers and their ratios. By the time of Euclid somebody, most likely a mathematician called Eudoxus, had worked out how to discuss magnitudes of any kind of size; this work forms the impressive Books V and VI of Euclid's *Elements*.

Accurate modern accounts of what we can reasonably say about Pythagoras can be found in the online *Stanford Encyclopedia of Philosophy* (Huffman, 2014) and in the book *The Cult of Pythagoras* (Martinez, 2012).

3.2 Euclid

We may know very little about what Pythagoras did, but we do have scholarly editions of the 13 books that compose Euclid's *Elements*. They give a rich picture of how Greek mathematicians put together the foundations of geometry and then built up a wide variety of results. Because Books VII, VIII and IX are about numbers, and are much more elementary than the books on either side, it is customary to say that in them Euclid collected earlier results (even, some say, Pythagorean ones). More precisely, the first six present a treatise on elementary plane geometry, Book X is on incommensurables and the final books, XI, XII and XIII, turn to the topic of solid geometry and culminate in the construction of the regular solids. Of the three on number theory, Book VII is concerned with various types of number – odd, even, square, prime, composite, etc. – and includes 'Euclid's Algorithm' for finding common factors. Book VIII is concerned with simple properties of squares and cubes, and contains some results on numbers in geometric progression. Book IX contains a number of theorems of special interest, the pearl being Euclid's celebrated proof of the existence of infinitely many primes, which we met in Book A, Chapter 2, Section 3.

Over one thousand editions of Euclid's masterpiece have appeared since the first printed version in 1482 and, to this day, very few books have been circulated and studied more widely. So it would be very pleasant if we could say something about the author of this bestseller, the *Elements*. But very little indeed is known about him. Most histories of mathematics say that he was born around 330 BCE, but even this date is disputed by the leading expert on Greek mathematics, Alexander Jones, who concludes his recent analysis of what Greek and Roman authors said about Euclid as follows (unpublished private communication):

> But the authentic Euclid is still hard to pin down. He does seem to have lived several decades later than we have been accustomed to say, and he was a more advanced and original mathematician than we would assume from the *Elements* and his other surviving works. But beyond this Euclid remains, as he was already in antiquity, an enigma, the most faceless of the great Hellenistic mathematicians.

3.3 Marin Mersenne

Mersenne (Figure 3.1) was a Franciscan friar, spending the main portion of his life in the Minim convent in Paris. From there he corresponded and became acquainted with many of the important mathematicians of that period, including Fermat. Indeed, Mersenne was regarded as a 'centre for correspondence' for mathematics at that time.

In 1644 Mersenne published the work for which he is best known, *Cogitata Physico-Mathematica*. In the preface to this book he made the claim that the only primes $p \leq 257$ for which $M_p = 2^p - 1$ is also prime are 2, 3, 5, 7, 13, 17, 19, 31, 67, 127 and 257. The last of these, 257, is the 55th prime and so this statement includes the claim that $2^p - 1$ is composite for the other 44 primes p that do not exceed 257.

Figure 3.1 Marin Mersenne (1588–1648)

How Mersenne reached his conclusions will never be known, but it is clear that with what was available at that time he could not have tested primality of each of the numbers by hand calculation. In fact it took 303 years and a lot of labour to check out the whole of Mersenne's claim. It is remarkable how close to the truth his claim was, for we now know that he made just five mistakes.

3.4 Pierre de Fermat

The seventeenth century saw three remarkable breakthroughs in mathematics. In the 1630s Descartes successfully introduced algebraic methods into the study of geometry, which made it possible for mathematicians to define and study large classes of curves for the first time. In the 1660s Newton and Leibniz independently invented the calculus. Less successful in its day was the work of Fermat on number theory, but it became one of the most important sources of the subject.

Fermat (Figure 3.2) was born near Toulouse, France, the son of a leather merchant. The year of his birth is uncertain, and he lived a quiet and uneventful life in and around his birthplace; from the age of 30 until shortly before his death in 1665 he was a lawyer and magistrate in the local parliament of Toulouse. His education in advanced mathematics seems to have begun during his time in Bordeaux in the late 1620s, and his first contributions were to the new analytic geometry that was rapidly dominated by Descartes. One offshoot of his work was that he discovered methods for finding tangents to curves and their maxima and minima well before Newton and Leibniz came on the scene.

Figure 3.2 Pierre de Fermat (1601?–1665)

Fermat seldom sought to publish his work and often circulated his ideas among a small group of mostly French mathematicians, and this was the case with his work on number theory. In 1657 he challenged the British mathematicians John Wallis and William Brouncker to show that there are always integer solutions to equations of the form $x^2 - Ay^2 = 1$, where A is positive and has no square factors – a challenge they eventually met – and we know from his choice of specific values of A, such as 109, that Fermat had discovered a general method, because the smallest values of x and y in this case are $x = 158070671986249$ and $y = 15140424455100$. These are not numbers you discover by trial and error.

Fermat's Last Theorem is the subject of Book C, Chapter 12.

One of Fermat's most famous 'theorems' is the 'Last Theorem', which was neither a theorem in his hands nor the last result he left unproved at his death. He owned a copy of Bachet's early seventeenth-century edition of the Greek mathematician Diophantus's *Arithmetica*, and in the margin by Book II Prop. 8, where Diophantus discussed the problem of dividing a given square number into two squares, Fermat wrote:

> On the other hand it is impossible to separate a cube into two cubes, or a biquadrate into two biquadrates, or generally *any power except a square into two powers with the same exponent*. I have discovered a truly marvellous proof of this, which however the margin is not large enough to contain.

The book in which he wrote this marginal note seems to be lost; the note itself was only published after his death, in the edition of his works made by his son Samuel in 1679. Since Fermat wrote nothing else on this subject that has survived, and as no one has ever found a proof with the means available to Fermat (as opposed to those of Andrew Wiles in 1995), the most likely thing is that Fermat simply made a mistake.

See Fermat (1679).

As for Fermat's genuinely last outstanding claim, it concerns the question of which integers can be the areas of right-angled triangles with sides that are rational numbers. Arabic writers knew of a triangle of this kind with area 5, which is now known to be the smallest possible value for the area, but the general problem remains unsolved.

Fermat came to this problem via another problem Diophantus had raised about triangles, in this case ones whose areas plus or minus a number are squares. Bachet, in his edition, remarked that if A is the area of a right-angled triangle with integer sides then there is number K such that $(2A)^2 - K^4$ is a square, and indeed if

$$\left(2pq(p^2 - q^2)\right)^2 + K^4 = s^2$$

then we may set

$$s = 2pq(p^2 + q^2),\ K = 2pq,$$

as you can readily check.

Fermat investigated this in the case when A is required to be a square, and was led to look for integer solutions to $x^4 + y^4$ being a square. He then showed that there is not even a solution in integers to the weaker equation $x^4 + y^4$ equalling a fourth power, which implies that there are no integer solutions to $x^4 + y^4$ equalling a square. In this case we do have Fermat's argument, which shows not only that he did tackle the $n = 4$ case of his 'Last Theorem' successfully but, more importantly, that he introduced into number theory a truly powerful new idea. He argued that one can prove certain equations do not have solutions in positive integers by arguing that *if* they do *then* they have a solution in smaller positive integers. This would give rise to an infinite descending sequence of positive integers, which is impossible. This method, known as Fermat's method of infinite descent, has proved invaluable to the present day.

See Book C, Chapter 12.

3.5 Leonhard Euler

Whilst Fermat's work on number theory had briefly revitalised the subject, he failed to interest many other distinguished mathematicians of his time in his work on numbers. The first to really appreciate Fermat's work was Leonhard Euler (Figure 3.3).

Euler was born in Basel, Switzerland. At the age of 13 he was sent to the University of Basel to study theology so that he might become a clergyman, like his father. However, at university Euler met the son of the distinguished mathematician Johann Bernoulli, who was the leading exponent of Leibniz's calculus, and who invited Euler to visit him regularly

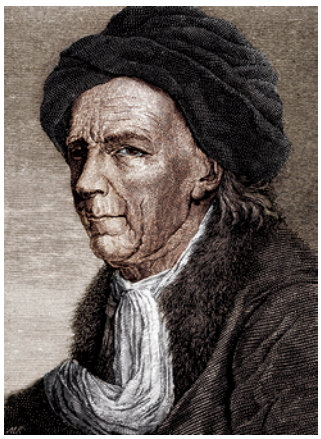

Figure 3.3 Leonhard Euler (1707–1783)

on Saturdays to discuss mathematics. Bernoulli recognised his talents at once, and went on over the years to praise him far more strongly and unreservedly than he ever praised anyone in his own highly talented mathematical family. Finally, in 1745, Bernoulli called Euler 'incomparable, the prince of mathematicians'.

Euler's career took off in 1727 with a call to the new Academy being set up in St Petersburg – the city Tsar Peter the Great had created as the face of modern Russia, in many ways the opposite of Moscow. The Tsar had decided to found an Academy of Science there in 1724, and when he died in 1725 his widow Catherine I took it over. Johann Bernoulli declined an invitation to move there, but his sons Niklaus II and Daniel I did go, and they and Christian Goldbach, the first permanent secretary of the Academy, lobbied for a position for Euler.

Euler's time there was generally a happy and productive one. His output was extraordinary. He had published his first paper (on a design for the arrangement of masts on a ship) at the age of 18, before going to St Petersburg – and before he had ever seen a ship. He went on to publish over 500 papers and many books during his lifetime. After his death, it took the St Petersburg Academy 47 years to publish all the manuscripts he had left behind. His writings covered the complete spectrum of mathematics including applications of calculus, algebra, coordinate and differential geometry, infinite series and number theory. He survived a life-threatening illness in 1735, but another illness that flared up in 1738 cost him the sight of his right eye. (He did not lose his sight through overwork, as is sometimes asserted.)

In 1740 Euler decided to leave St Petersburg for Berlin, to manage the Academy (later, the Académie Royale des Sciences et des Belles-Lettres) that the Prussian Emperor Frederick the Great was establishing in Berlin with Maupertuis as its first President. He stayed until he came to realise that he had accomplished all he could there, leaving Berlin on 1 June 1766 and returning to St Petersburg on 28 July, to a triumphant reception. There he stayed, productive as ever even though the sight in his remaining eye was deteriorating, until on the afternoon of 18 September 1783 he was felled by a massive stroke. In the words of the French mathematician the Marquis de Condorcet, at around 11 o'clock that night 'The great Euler ceased to calculate and to live'.

Euler's *Collected Works* runs to 78 volumes, and undoubtedly one of the results of his prolific publication at many levels was to make mathematics accessible to a wider audience than ever before. Four of these volumes record his explorations of number theory, and more results are to be found elsewhere. His extensive calculations with binary quadratic forms came close to discovering the law of quadratic reciprocity. The zeta function is one of his discoveries, as is the remarkable result that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

Euler brought the subject out of the backwater to which generations had consigned Fermat's work and made it worthy of the attentions of Lagrange and, in due course, Gauss.

3.6 Joseph-Louis Lagrange

Lagrange (Figure 3.4) was born in Turin to a French father and Italian mother, but he spent the most productive years of his life in Berlin and then Paris. His father intended him to become a lawyer, but Lagrange and others soon realised that he had a particular talent for mathematics, and he became appointed a professor at the Royal Artillery School in Turin at the age of 19.

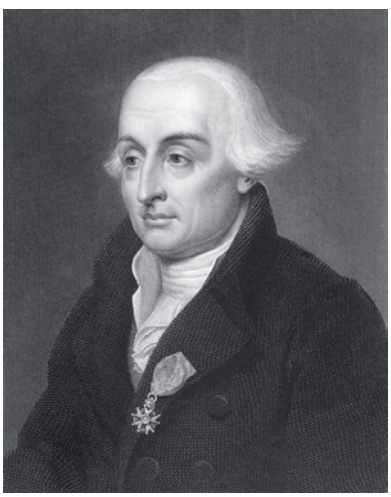

Figure 3.4 Joseph-Louis Lagrange (1736–1813)

Lagrange sent many of his earliest discoveries to Euler, who at the time worked at the Academy of Sciences in Berlin, and although Euler recognised a great talent in the young Lagrange they never met. In 1766, when he was due to leave the Berlin Academy for St Petersburg, Euler arranged for Lagrange to be installed as his replacement. A message from Frederick the Great invited 'the greatest mathematician in Europe' to live near 'the greatest king in Europe'. Lagrange accepted, and worked in Berlin for 20 years until, on the death of Frederick in 1786, he was head-hunted by the French and moved to Paris the next year.

For the next six years Lagrange contributed nothing to mathematics. Strangely, it was the turmoil of the French Revolution that revitalised his interest in the subject. Following the abolition of the old universities, Lagrange became a professor at the newly founded École Polytechnique. In this period he published his attempt to establish the calculus on a purely algebraic footing, the *Théorie des fonctions analytiques* (1797).

Although Lagrange's main researches were in the domains of analytical mechanics, theoretical astronomy, calculus and function theory, he was also a profound number theorist. In 1770 he was the first to prove Fermat's claim that every positive integer can be expressed as the sum of four squares, and he often turned Euler's provocative speculations into proven theorems. He was the first to propose that a binary quadratic form $ax^2 + bxy + cy^2$ can be reduced to a form in a finite list of such forms with the same value of $b^2 - 4ac$; Lagrange's reduction theory was a significant step in bringing coherence to the findings of Fermat and Euler, and Euler was particularly pleased by it.

3.7 Adrien-Marie Legendre

Legendre (Figure 3.5) was born in Toulouse and educated at the Collège Mazarin in Paris. Throughout a distinguished mathematical career he contributed to significant advances in many fields. He was mainly an analyst, who wrote a three-volume comprehensive treatment of integral calculus, followed by a further three-volume work on elliptic integrals. In these works he introduced numerous tools of analysis still used by applied mathematicians, including the beta and gamma functions and the

Legendre polynomials, which are used in solving certain differential equations. He also wrote an introduction to geometry, his *Éléments de géométrie*, that was influential in swinging geometry teaching back to the formal style of Euclid and remained in use in France for much of the nineteenth century, but which was marred in successive editions by flawed attempts at defending the parallel postulate and therefore denying the possibility of non-Euclidean geometry.

Figure 3.5 Adrien-Marie Legendre (1752–1833), shown on the left with Fourier beside him; this splendid cartoon seems to be the only authentic picture of Legendre in existence!

Legendre was not a great innovator in number theory but his writings did point mathematicians of the nineteenth century towards fruitful areas for investigation. In 1798 he published a two-volume work, *Essai sur la théorie des nombres*, probably the first treatise devoted solely to number theory; an expanded edition, *Théorie des nombres*, was printed in 1830. Amongst many other results, he completed an argument begun by the young Dirichlet in 1825 that proved the equation $x^5 + y^5 = z^5$ has no solution in the natural numbers – this is a particular case of Fermat's Last Theorem. On the question of distribution of primes, he conjectured that $\pi(n)$ approaches $\dfrac{n}{\log(n) - 1.08366}$ as n increases, a result that is very close to being correct.

In his *Essai* Legendre included an account, together with an attempted proof, of the Law of Quadratic Reciprocity. The relationship given in the Law was first conjectured by Euler in 1783, on the basis of considerable numerical evidence. Two years later Legendre offered the first proof. This proof, like the later one published in the *Essai*, was seriously flawed. It broke into eight cases, in one of which Legendre relied on the plausible

claim that every arithmetic progression $\{a + bk : k = 1, 2, \ldots\}$ with a and b relatively prime contains infinitely many primes. However, he could not prove this claim. Worse, in another of the eight cases Legendre's reasoning was circular. Gauss, in his *Disquisitiones Arithmeticae* (1801), acknowledged that, before him, Euler had almost discovered the result and that Legendre knew it and had proved it in various special cases, but he attributed the Law to himself on the grounds that his two proofs were the first valid ones. This Legendre was unwilling to accept, but the result about primes in arithmetic progressions was only to be established by Dirichlet in 1837, and the circularity in Legendre's line of reasoning has not been removed to this day.

3.8 Joseph Fourier

Joseph Fourier (Figure 3.6) was born in Auxerre in France and orphaned at the age of 9, at which point he was placed in the town's military school. There he discovered both a passion for mathematics and the strength of his calling to civic and military duty. The French Revolution found him back in his home town as a school teacher. He became involved in politics and courageously opposed the Terror, the bloodthirsty events of 1794 that saw many people dragged without trial to their deaths on the guillotine, and he was arrested. A personal appeal to Robespierre failed, but when Robespierre was himself executed on 28 July 1794, Fourier was released and went to the École Normale. After a year he was appointed as an assistant lecturer at the École Polytechnique, working under Lagrange and Monge, the French geometer who was the founding father of the École Polytechnique. In 1798 Monge, by now a prominent supporter of Napoleon, selected him to go on the French expedition to Egypt. This was a French attempt to exclude the British from the eastern Mediterranean, but it had a colonial, anthropological side, and French engineers made impressive surveys of Egyptian pyramids and other remains.

Figure 3.6 Joseph Fourier (1768–1830)

With the defeat of the French by the British in 1801, Fourier returned to France, hoping to resume his work at the École Polytechnique, but Napoleon had noticed Fourier's impressive organisational talents and sent him instead to be the Prefect (or Governor) of the Department of Isère, which at that time extended from Grenoble to the French border. He did this so well that Napoleon made him a baron in 1808. In 1809 he finished his contribution to the *Description d'Egypt*, a massive account of ancient Egypt based on the 1798 surveys, which marks the start of the celebration of ancient Egypt in the modern world. (Cleopatra's and other needles began their journeys at this time, as did more and more fanciful statements about ancient wisdom, secret knowledge and so forth, none of which can be held against Fourier.)

The eventual defeat and final exile of Napoleon were the lowest point of Fourier's life, but in 1816 a former colleague on the Egypt expedition rescued him and gave him a job as Director of the Bureau of Statistics for the Department of the Seine, a position that left him good time for research. He now suffered for his former associations with Napoleon, but

after more protests, he was appointed to the reformed Academy of Sciences, rising to become its permanent secretary in 1822. He was elected to the prestigious Académie Française in 1827 (by which time he was 59). He died in 1830 as the result of complications from an illness caught in Egypt.

Fourier's great work on heat diffusion, the *Théorie analytique de la chaleur*, has been hailed as one of the few major contributions to both pure and applied mathematics. It was written in several stages. The first version was finished in 1807 and submitted to the Paris Academy of Sciences, but Lagrange blocked publication because Fourier's ideas contradicted some that Lagrange had published in the 1750s. In 1810 the Academy of Sciences announced a prize competition on heat diffusion. Fourier revised his memoir and won the competition, but the memoir was still criticised for a lack of rigour and generality. Fourier thought the criticism unfair, but revised it for publication, and the resulting book came out in 1822, when his personal circumstances had much improved.

3.9 Carl Friedrich Gauss

Figure 3.7 Carl Friedrich Gauss (1777–1855)

Gauss (Figure 3.7) was unquestionably one of the greatest mathematicians of all time, and it is clear from his work that he possessed an ability, rare among mathematicians, for doing elaborate arithmetical calculations in his head. He himself told the story of being a young child in school and surprising his teacher by summing an arithmetic progression immediately the question was posed. But if that story survives in a reputable source, many of the details in histories of mathematics have been added for effect, such as the brutality of the teacher and the actual arithmetic progression (commonly said to be $1 + 2 + \cdots + 100$).

However, we do not need to exaggerate Gauss's merits. He made his name in 1801, the year he turned 24, for two remarkable achievements. The first was in astronomy. On 1 January 1801 the Italian astronomer G. Piazzi had discovered what we now call the asteroid Ceres, the first new body in the solar system since Herschel's discovery of Uranus 20 years before. But it was already close to the sun, and Piazzi was only able to observe it for 42 days before it vanished behind the sun; thus his observations, with all the observational errors they necessarily contained, did not make it easy to determine its orbit. Piazzi published his data in June, and invited Europe's astronomers to predict where it would next be visible. Gauss took up the challenge, and although his prediction was quite widely separated from the others it was clear by the end of 1801 that he was correct. He had succeeded where the others failed because he had been willing to take on two demanding tasks. He allowed that Ceres might have an elliptical orbit rather than a circular one, and he invented the method of least squares to handle Piazzi's data. Such was the excitement caused by the new 'planet' that Gauss briefly tasted European fame – and also controversy, because he did not reveal his methods, so when the distinguished French mathematician Legendre discovered the method of least squares a few years later and published it, he refused to accept Gauss's claims to priority.

Gauss's second great discovery, and the one of more interest to us, was his reformulation of number theory, which he presented in his book *Disquisitiones Arithmeticae* (*Arithmetical Researches*) in 1801. It opens with an account of modular arithmetic, in which he was the first to prove that the multiplicative group of non-zero residues modulo a prime p is cyclic; a generator is called a primitive root mod p. He then showed how modular arithmetic enables one to take up the theory of quadratic forms and solve most of the outstanding problems in the subject. Since this was a topic that Euler, Lagrange and Legendre had repeatedly tackled, Gauss was aiming high indeed.

Some of the better-known results by Gauss's day were as follows.

1. Every odd number of the form $x^2 + y^2$ is a product of prime numbers of the form $4n + 1$ and squares.
2. Every prime number of the form $4n + 1$ is a sum of two squares (i.e. is of the form $x^2 + y^2$).
3. Every divisor of a number of the form $x^2 + 2y^2$ is of that form, and every divisor of a number of the form $x^2 + 3y^2$ is again of that form ...
4. ... but divisors of numbers of the form $x^2 + 5y^2$ are either of that form or of the form $2x^2 + 2xy + 3y^2$ (consider $1^2 + 5 \times 2^2 = 21 = 3 \times 7$).
5. For a given quadratic form $ax^2 + 2bxy + cy^2$, where a, b, c have no common factors, and a given integer m, if there are integers x_0 and y_0 such that $ax_0^2 + 2bx_0y_0 + cy_0^2 = m$ then $b^2 - ac$ is a square mod m. (This is why modular arithmetic enters the theory.) We say that in this case the quadratic form represents the number m.

Gauss took the subject and gave a series of tests that one could apply to find the possible quadratic forms that represent a given number. His tests are easy to apply, his analysis greatly extends and deepens what was known, and his proofs are extremely difficult. This is easy for us to see. In essence he found a way of applying ideas from group theory to the study of quadratic forms, and not just group theory but the study of group characters (maps from a group to the group of non-zero complex numbers). Indeed, the word 'character' in this context was introduced by him. But – and it is a very important 'but' – group theory did not exist in his day, and was not to do so for a generation!

Among the other important results in the *Disquisitiones Arithmeticae* are the first two proofs of the Law of Quadratic Reciprocity, which makes it very easy to handle the question of when one prime is a square modulo another. This occasioned another dispute with Legendre, who had previously claimed a proof of this Law, but one that Gauss had to observe was seriously flawed.

The subsection of the *Disquisitiones Arithmeticae* that proved most popular goes by the unwieldy name of cyclotomy, or circle division. It concerns what can be said about the roots of unity. Because

$$x^n - 1 = (x - 1)(x^{n-1} + x^{n-2} + \cdots + x + 1),$$

the roots of the equation $x^n - 1 = 0$ other than 1 satisfy the equation

$$x^{n-1} + x^{n-2} + \cdots + x + 1 = 0.$$

Even before going to university, Gauss had discovered that this equation can usually be solved by solving a sequence of equations of lower degree. For example, when $n = 17$, the equation for the non-trivial 17th roots of unity is

$$x^{16} + x^{15} + \cdots + x + 1 = 0.$$

Gauss showed that it is possible to find a quadratic equation the roots of which allow one to factorise this equation into two factors of degree 8, and then to repeat this trick by solving another quadratic, the roots of which allow one to factorise each of these polynomial factors of degree 8 into two factors of degree 4, and so on. In the end all the roots of unity are found by solving nothing worse than quadratics, and because solving a quadratic equation is equivalent to carrying out a construction by straightedge and circle, Gauss had found the first regular polygon to be constructible that way since those known to the Greeks. The Greeks had given constructions for the regular or equilateral triangle, the square, the regular pentagon, and all the regular figures constructible from these by bisecting their sides, but the regular 17-gon was new. Gauss went on to claim, without giving a proof, that a regular N-gon can be constructed with compass and straightedge if, and only if, $N = 2^r p_1 p_2 \cdots p_k$, where the primes p_i are distinct Fermat numbers and $r \geq 0$. (A Fermat prime is a prime number of the form $2^{2^n} + 1$, such as 3, 5 and 17.)

Gauss retained a lifelong interest in astronomy, and did a lot of work on the orbits of the first few asteroids to be discovered. He was employed in the 1820s to conduct a survey of the state of Hanover, where he lived (the Hanoverian monarchs of England are cousins of the German Hanoverians), and this led him to invent the heliotrope and to discover the concept of the intrinsic curvature of a surface, which decisively reformulated the subject of differential geometry. In the 1830s he did important work on potential theory, which he applied to electricity and magnetism. He also had original things to say about differential equations and complex analysis, and left many profound ideas unpublished at his death. But it seems that number theory was his favourite subject, because of the contrast between the ease with which results can be conjectured and the frequent difficulty found in proving them. Moreover, the proofs often made unexpected and deep connections with other branches of mathematics. To that end, for example, he published a total of six proofs of quadratic reciprocity, and left a further two unpublished.

As a young man, Gauss kept a diary in which to record his discoveries. It did not come to light until some forty years after his death, and its brief statements of 146 results amply display his genius. The first entry concerns cyclotomy, and runs ‘The principles upon which the division of the circle depend, and geometrical divisibility of the same into seventeen parts, etc.’ (30 March 1796, when Gauss was still 18); the sixteenth (for 27 June 1796) concerns the second proof of the Law of Quadratic

Reciprocity and says 'A new proof of the golden theorem all at once, from scratch, different, and not a little elegant'.

3.10 Peter Gustav Lejeune Dirichlet

Dirichlet (Figure 3.8) is one of the most important mathematicians of the nineteenth century. A German despite the French appearance of his name, he was born in Düren in 1805. At the age of 17 he took himself off to Paris to study mathematics, for the level in Germany was very low and Paris was then by far the leading place for mathematics in the world. He made many good social contacts; through them he met Fourier, and acquired from him a strong interest in trigonometric series and mathematical physics. His own interests in mathematics, however, initially concentrated on number theory. At the age of 20 he presented a memoir to the French Academy of Sciences that almost completely resolved the first new case of Fermat's Last Theorem since Euler had dealt with cubes, by showing that many cases of the equation $x^5 + y^5 = z^5$ cannot be solved in integers. The remaining cases were solved soon afterwards by Legendre, but Legendre's publication unduly minimised the extent to which the credit was due to Dirichlet.

Figure 3.8 Peter Gustav Lejeune Dirichlet (1805–1859)

Dirichlet returned to Germany in 1826, and in 1828 was appointed a Professor of Mathematics at the new University of Berlin, founded by Alexander von Humboldt in 1809. In 1831 he married Rebecca Mendelssohn-Bartholdy, the sister of the famous composer; Dirichlet was a good pianist in his own right. It would take us too far afield to discuss Dirichlet's many contributions to number theory, but he became the person who explained Gauss's work to the rest of the world, thus helping to make number theory into a major branch of pure mathematics, and almost the leading branch in Germany.

Equally important was Dirichlet's attitude to mathematical rigour, captured marvellously in the following remark of his lifelong friend Carl Jacobi:

> If Gauss says he has proved something, it seems very probable to me; if Cauchy says so, it is about as likely as not; if Dirichlet says so, it is certain. I would gladly not get involved in such delicacies.

Jacobi to A. von Humboldt, 1846, quoted in H. Pieper's 'Gegen die Schmach des Belagerungszustands' (1980, p. 23). The title means 'Against the ignominy of a state of emergency'.

Dirichlet may be called the man who brought rigorous mathematics to Germany, if not indeed to the mathematical world, and he is particularly remembered for his work on potential theory and his work on the convergence of Fourier series, which could serve as a textbook example of how to do epsilon-delta analysis to this day.

When Gauss died in 1855 Dirichlet, who was then 50, moved to Göttingen as his successor. But his hopes of a new and active phase of life were cut short by a heart attack when he was 58, and within a year he was dead.

3.11 Bernard Riemann

Riemann (Figure 3.9) was the archetype of the shy mathematician, close to his family and awkward outside it. We learn from a recent biography of him that his interests were almost exclusively mathematics, physics and philosophy. He was devout in his religion, and had initially intended to study theology at Göttingen University, in accordance with his father's wishes, before his remarkable ability at mathematics led him to switch subjects. But even though Gauss was a professor there, the level of mathematical education in Göttingen was not particularly high, and in accordance with the German freedom to study anywhere Riemann left Göttingen for Berlin in 1847, where he spent two years learning from Dirichlet about potential theory and partial differential equations, number theory and theory of integration.

See Laugwitz (2000).

Figure 3.9 Bernard Riemann (1826–1866)

But Berlin was a bustling capital, and in other ways Göttingen was the ideal choice for Riemann. He returned there in 1849 and attended Wilhelm Weber's lectures on mathematical physics, and for a while he devoted himself to studies in physics and philosophy of science. This influenced his way of dealing with geometry. In Göttingen he soon met Richard Dedekind, who was five years younger than him, and who was also a truly conceptual thinker, but unlike Riemann one who was to move slowly and steadily throughout a long life.

In 1855 matters improved greatly for Riemann. Gauss had for some years been a distant but real presence in Göttingen, but after his death in 1855 Dirichlet was called as the senior professor. Dirichlet was the ideal abstract thinker to guide Riemann, and the young man remained very grateful to him. His influence is visible in Riemann's work, not just in the published papers but in the lectures he began to give.

When Dirichlet died in 1859, Riemann was quickly appointed his successor (the more charismatic figure of Dedekind being by then a professor in Zürich). He still attracted only a few students, and one wonders how matters might have developed had he not contracted pleurisy in 1862. This led to a permanent weakening of his lungs, and he was advised to spend as much time as possible in the South. He accordingly spent as much time as he could in Italy, but his health deteriorated and on 20 July 1866 he died near Lake Maggiore, where he is buried. He left behind a number of published papers, several more in a good enough state to be published, and yet more that could be edited and printed in the first edition of his *Werke* (edited by Dedekind and Heinrich Weber in 1876).

See Riemann (1851, 1854).

3.12 Georg Cantor

Cantor (Figure 3.10) was born in St Petersburg, Russia, but his family soon moved to Germany. He studied mathematics in Berlin, where he was greatly influenced by Weierstrass, Kummer and Kronecker, the triumvirate who dominated mathematics there in the second half of the nineteenth century. After writing his doctorate he obtained a teaching position at the University of Halle, where he remained for the rest of his life, becoming a

full Professor there in 1879, although he always hoped to be offered a position back in Berlin.

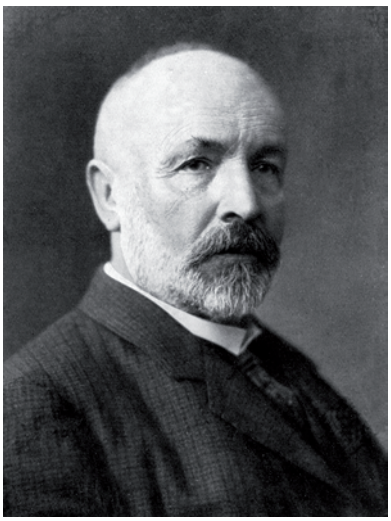

Figure 3.10 Georg Cantor (1845–1918)

Cantor's first work was on the anomalous behaviour of Fourier series and this led him to develop important ideas in point-set topology. Only several years later did he realise that this work opened the door to a new theory of infinite sets in which it was possible to construct sets of different orders of infinity. In 1873 he gave the first proof that the real numbers are not countable, and thereafter his research was focused on the task of creating a rigorous theory of transfinite sets. Opinions of this work varied. Weierstrass was impressed by it, and for a time the Swedish mathematician and energetic journal editor Gösta Mittag-Leffler published it in his journal *Acta Mathematica*; when Cantor thought Mittag-Leffler was losing interest in it he switched to another recently founded journal, Felix Klein's *Mathematische Annalen*. But others opposed it, notably Kronecker, and the big Berlin-based mathematical journal, Crelle's *Journal für Mathematik*, was closed to him. Hermite, the influential French mathematician, found Cantor's work pointless, but French mathematicians of the next generation – such as Poincaré – found the topology useful.

By 1900, Cantor's work had become caught up in heated debates about the foundations of mathematics. It proved easy to find paradoxes in the naive theory of sets, and although Cantor regarded these as harmless and inevitable when finite human minds deal with the concept of infinity, others realised that it was perhaps better to build up a less naive but more rigorous account of set theory. As a result, this became a task for mathematicians in the twentieth century. David Hilbert made Cantor's continuum hypothesis the very first of the problems he posed the international community in 1900. It addresses the size of the set of all real numbers, which is denoted c. The smallest infinite set is the integers, $\mathbb{Z}$, and it has a size denoted $\aleph_0$. Cantor proved that the set of all subsets of $\mathbb{Z}$ is strictly bigger than $\mathbb{Z}$ and said it has size $\aleph_1$. The set of all subsets of this set is bigger again and is said to have size $\aleph_2$, and so on. The question arises of where c fits in to all of this, and Cantor's hypothesis was that $c = \aleph_1$. He spent many years trying to prove it, unsuccessfully; it is now known to be independent of the usual axioms for set theory that were developed after Cantor's death.

Cantor was the first President of the German Mathematical Society when it was established in 1890, but he began to suffer prolonged disabling bouts of depression and was frequently hospitalised. He eventually died in the psychiatric clinic of Halle University.

3.13 Felix Hausdorff

Hausdorff (Figure 3.11) was born in Breslau, Germany (now Wroclaw, Poland) in 1868. He became attracted to philosophy when a student; under the name of Paul Mongré, he went on to write two books on philosophy in an upbeat, Nietzschean spirit, as well as a successful play, before turning to mathematics in his forties. His philosophy led him to wonder what, if anything, could be provably true about the world we live in, and he decided

Figure 3.11 Felix Hausdorff (1868–1942)

that the ideas that time is linear, continuous, single-valued, does not run backwards and proceeds uniformly, and that space is three-dimensional, continuous, uniform and Euclidean, are incapable of proof.

To support his views Hausdorff turned to mathematics, especially Cantor's new theory of infinite sets, and after 1911 his interests became almost exclusively mathematical. His *Grundzüge der Mengenlehre* (1914) marked a significant advance in both set theory and topology. In 1915 Hausdorff published what became known as his paradox, which shows that on any definition of measure that satisfies Lebesgue's four axioms there will be non-measurable sets.

For many years after the First World War he taught mathematics at the University of Bonn, but the Nazi accession to power in 1933 led to years of persecution, because Hausdorff was a Jew. With his wife, he committed suicide in the town of Endenich in January 1942 to avoid deportation to a concentration camp, bravely writing the note 'Und vielleicht ist Endenich auch das Ende nicht' ('and perhaps Endenich is also not the end').

References

Cardano, G. (1663) 'Vita Ludovici Ferrarii Bononiensis', in *Opera Omnia*, vol. 9, Lyons.

Cardano, G. (1543) *Ars Magna* (trans. T.R. Witmer), MIT Press (this edition 1968).

Cauchy, A.L. (1815) 'Mémoire sur le nombre des valeurs qu'une fonction peut acquérir, [etc.]', *Journal de l'École Polytechnique*, vol. 10, pp. 1–27, in *Oeuvres Complètes*, series 2, vol. 1, pp. 64–90.

Cauchy, A.L. (1844) 'Mémoire sur les arrangements que l'on peut former avec des letters données, [etc.]', *Exercises d'analyse et de physique mathématique*, vol. 3, pp. 151–252, in *Oeuvres Complètes*, series 2, vol. 13, pp. 171–282.

Dirichlet, P.G.L. (1876) *Vorlesungen über die im ungekehrten Verhältniss des Quadrats der Entfernung wirkenden Kräfte*, Leibzig, Teubner.

Fermat, P. (1679) *Varia Opera Mathematica*, Toulouse.

Fourier, J. (1822) *Théorie analytique de la chaleur*, Paris, Firmin Didot; English translation *The Analytical Theory of Heat*, Cambridge University Press, 1878 (reprinted Dover and New York, 1950).

Fréchet, M. (1906) 'Sur quelques points du calcul fonctionnel', *Rendiconti del Circolo Matematico di Palermo*, vol. 22, pp. 1–74.

Gauss, C.F. (1801) *Disquisitiones Arithmeticae* (trans. W.C. Waterhouse and A.A. Clarke), New York, Springer (revised edition 1986).

Hales, T.C. (2007) 'Jordan's proof of the Jordan curve theorem', *Studies in Logic, Grammar and Rhetoric*, vol. 10, no. 23, pp. 45–60.

Harnack, A. (1887) *Die Grundlagen der Theorie des logarithmischen Potentiales und der eindeutigen Potentialfunktion in der Ebene*, Liepzig, Teubner.

Hausdorff, F. (1914) *Grundzüge der Mengenlehre*, Leipzig, Veit.

Huffman, C. (2014) 'Pythagoras', in Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy* [Online], Spring 2014 edn. Available at http://plato.stanford.edu/archives/spr2014/entries/pythagoras (Accessed 2 May 2014).

Jordan, C. (1870) *Traité des substitutions et des équations algébriques*, Paris, Gauthier-Villars.

Jordan, C. (1887) *Cours d'analyse de l'École Polytechnique*, vol. 3, Paris, Gauthier–Villars.

Lützen, J. (2009) 'Why was Wantzel overlooked for a century? The changing importance of an impossibility result', *Historia Mathematica*, vol. 36, pp. 374–94.

Lagrange, J.-L. (1797) *Théorie des fonctions analytiques*, Paris, Imprimerie de la République; 2nd edn Paris, Courcier, 1813 in *Oeuvres Complètes*, vol. 9.

Lagrange, J.-L. (1770) 'Réflexions sur la résolution algébrique des équations', *Nouveaux mémoires de l'Académie royale des sciences et belles-lettres de Berlin*, in *Oeuvres Complètes*, vol. 3, pp. 205–421.

Laugwitz, D. (2000) *Brenhard Riemann*, Berlin, Birkhäuser.

Legendre, A.-M. (1798) *Essai sur la théorie des nombres*, Paris, Courcier; 2nd edn 1808.

Legendre, A.-M. (1830) *Théorie des nombres*, 3rd edn, Paris, Firmin-Didot.

Martinez, A.A. (2012) *The Cult of Pythagoras: Math and Myths*, University of Pittsburgh Press.

Morley, H. (1854) *The Life of Girolamo Cardano, of Milan, Physician*, vol. 1, London, Chapman and Hall.

Neumann, P.M. (2011) *The Mathematical Writings of Évariste Galois*, Heritage of European Mathematics series, Zürich, European Mathematical Society.

Pieper, H. (1980) 'Gegen die Schmach des Belagerungszustands', *Spectrum, Monatsbericht der AdW der DDR*, vol. 1, pp. 22–4.

Riemann, B. (1851) 'Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse', in *Werke*, pp. 35–80; this edition 1990, ed. R. Narasimhan, Springer.

Riemann, B. (1854) 'Über die Darstellbarkeit einer Function durch einer trigonometrische Reihe', *Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen*, vol. 13, pp. 87–132, in *Werke*, pp. 227–71; this edition 1990, ed. R. Narasimhan, Springer.

Robson, E. (2007) 'Mesopotamian mathematics', in Katz, V. (ed.) *The Mathematics of Egypt, Mesopotamia, China, India, and Islam: a Sourcebook*, Princeton University Press, pp. 58–186.

Rothman, T. (1989) *Science à la Mode*, Princeton University Press.

Schwarz, H.A. (1870) 'Zur integration der partiellen Differentialgleichung $\Delta u = 0$', *Monatsberichte der Königlich preussischen Akademie der Wissenschaften zu Berlin*, pp. 767–95, in *Gesammelte mathematische Abhandlungen*, vol. 2, pp. 144–71.

Serret, J.-A. (1849) *Cours d'algèbre supèrieure*, 1st edn, Paris, Gauthier-Villars; 2nd edn 1854, 3rd edn 1866.

Smith, H.J.S. (1875) 'On the integration of discontinuous functions', *Proceedings of the London Mathematical Society*, vol. 6, pp. 140–53, in *Collected Mathematical Papers*, vol. 1, pp. 86–100, Oxford University Press (this edition 1965).

Steinitz, E. (1910) 'Algebraische Theorie der Körper', *Journal für Mathematik*, vol. 137, pp. 167–309; reprinted 1930 with additional material by R. Baer and H. Hasse, Leipzig, de Gruyter.

Tignol, J.-P. (2001) *Galois' Theory of Algebraic Equations*, World Scientific.

Veblen, O. (1905) 'Theory on plane curves in non-metrical analysis situs', *Transactions of the American Mathematical Society*, vol. 6, pp. 83–98.

Wantzel, P.L. (1837) 'Recherches sur les moyens de reconnaitre si un problème de géométrie peut se résoudre avec la règle et le compas', *Journal des mathématiques pures et appliquées*, vol. 2, pp. 366–72.

Acknowledgements

Grateful acknowledgement is made to the following sources.

Figure 1.1: Wellcome Library, London. This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/4.0/.

Figure 1.2: Municipal Archives of Trondheim/Wikimedia.org. This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/2.0/.

Figure 1.3: Wellcome Library, London. This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/4.0/.

Figure 2.7: Pruneau/Wikimedia.org. This file is licensed under the Creative Commons Attribution-Share Alike Licence http://creativecommons.org/licenses/by-sa/3.0/.

Figure 2.10: MacTutor History of Mathematics archive.

Figure 3.2: © Bettmann/CORBIS.

Figure 3.3: © INTERFOTO/Alamy.

Figure 3.4: © iStock.com/GeorgiosArt.

Figure 3.5: Licensed with permission from Bibliothque de L'Institut de France/RMN-Grand Palais.

Figure 3.6: Stock Montage/Getty Images.

Figure 3.7: © Heritage Image Partnership Ltd/Alamy.

Figure 3.10: © INTERFOTO/Alamy.

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.